A Spectral Theory of Token Communication

Operators, Amplification and the Architecture of Transformers

Neil D. B. Bruce
University of Guelph

Version 0.1 · 2026-10-22
https://doi.org/10.5281/zenodo.23541723

About this edition

Version. 0.1, 2026-10-22. Later versions revise the text as the work develops, and every version keeps the same concept DOI. This version: https://doi.org/10.5281/zenodo.23541723.
Licence. The text is released under the Creative Commons Attribution 4.0 International licence (CC BY 4.0).
Code and records. The records behind the book's numbers, the code that produced them and the formal proofs in Lean follow in a later version: each source paper's package is archived as its own record, cross-linked to this book (Appendix D).
Use of AI tools. AI tools were used, under the author's direction, for several kinds of work: software development and debugging, mathematical derivation and checking, data analysis, literature search and citation checking, and drafting and editing of the text. The author reviewed and verified all AI-assisted content to the extent possible and takes responsibility for the book.
How to cite. Neil D. B. Bruce. *A Spectral Theory of Token Communication: Operators, Amplification and the Architecture of Transformers*. Version 0.1, 2026-10-22. https://doi.org/10.5281/zenodo.23541723.

Foreword

All of the work in this monograph grew out of a small set of intertwined questions. How did we land on transformers? Why do they seem to work so well, and for so many things? Why have they become a fixed point, or at least a strong attractor, in the space of architectures? And why haven't we moved on to the next big thing, and what might the alternatives look like?

I wouldn't claim that this work answers any of these questions. Yet it has reached a natural stopping point. The answers so far are incomplete, but weighing my satisfaction with them against the focus, stamina and resources it would take to go further, I think it is time to pause. As the work went on, the program kept expanding: each answer raised new questions, and each new question demanded another experiment. At some point convergence had to be forced, and this monograph is the result.

Version 0.1. This version of the monograph is numbered 0.1, and the numbering is intentional. It is an early release, meant to let others build on whatever parts of the foundation prove valuable. It does not yet include all of the supporting provenance (the code, the records and the formal proofs), which will appear with version 1.0, once the material is fully organized and ready for release. What is here will continue to be refined, verified and documented, and doing that properly will take time.

How to read it. I take responsibility for this work, and parts of it may well turn out to be wrong, or emptier than they look. For now, please treat it with a healthy degree of skepticism. Don't trust a claim because it looks right on the surface, or because the figure a chapter leads with is attractive and the work therefore seems well vetted. Each claim of Chapter 0 carries a label that says what kind of support it has, such as proved, measured, partial or pending (Reading guide, R.2), and each chapter closes with what holds in trained networks. I invite you to scrutinize anything that appears here, to stress-test it, and to tell me where an argument is circular or vacuous, and where something is utter nonsense, incomprehensible, or even madness.

How it was made. There is much discussion of fast science and slow science. This work is not one or the other, and it is not a fast part sitting beside a slow part: all of it is a mix of the two. Much of the program took eight months, though a few of its seeds sat on my hard drive for a year or more before that. But those eight months were amplified and produced far more than one person could have produced in that time without AI tools, which figured in almost every part of the work, from carrying it out to communicating it. Still, the path travelled was mine, from the program's structure and framing to the specific cases it examined. The book's edition page carries a generic statement of how AI was used. I plan to replace it with a detailed account, not just of which results leaned most on AI tools, but of how information flows through a project like this and what the parts of the overall process look like. In some ways, the process itself feels like the system the work characterizes.

I have reviewed most of the details closely, checked them numerically and, in places, verified them formally in Lean, and they may still contain errors. I'm happy to admit that, despite my efforts, the mathematics behind some of the results the program aimed to establish is beyond my

reach, at least given the time it would take to digest every piece. In most cases the nature of the relationships established aligns precisely with what was sought, but the details and the edge cases turn out to be messier than one might expect.

What kind of work this is. It isn't clear that all the time dedicated to this project was well spent, but I believe that the concepts the work exposes offer ideas about what kind of machine is necessary for the computation a transformer carries out. Those concepts reside at several levels of abstraction from transformers. Abstraction is necessary in many places to constrain the system enough to make specific predictions, and some of those predictions carry further than others. Characterizing a system at several levels is likely to invite critics, and the invitation is welcome: one aim of the work is to stimulate discourse at the level of scientific understanding, rather than to add an appendage to a model, claim victory on a benchmark and stop there. The project also evolved from working and thinking at these levels, and connecting its concepts to the mediums available for communication isn't successful in every case. If the fuzzy nature of the science in this work is a concern, it may be more satisfying to consider the collection of constraints as much art as science. In the end, if this work inspires any ideas that turn out to be worthwhile, it has achieved its goal.

Contents

About this edition i

Foreword ii

0 The Transformer Abides 1
- 0.1 The question . . . 1
- 0.2 What the usual answers explain . . . 5
  - 0.2.1 Why this lineage won: hardware, scale and history . . . 5
  - 0.2.2 Why this design: the properties that matter . . . 5
  - 0.2.3 Why it stays: how the field behaves . . . 6
  - 0.2.4 What they leave open . . . 6
  - 0.2.5 Attention's strong bias is averaging . . . 8
- 0.3 A sweet spot of conditions . . . 8
  - 0.3.1 The object: averaging, rebuilt from its input and iterated . . . 8
  - 0.3.2 What fails without each of the six conditions . . . 9
  - 0.3.3 The sweet spot . . . 12
  - 0.3.4 Why the neighbourhood is flat . . . 16
  - 0.3.5 Do designs jump rather than drift? . . . 17
  - 0.3.6 Adopted changes in retrospect . . . 17
  - 0.3.7 Where the design can still move . . . 18
  - 0.3.8 How strong this case is . . . 21
- 0.4 What the question uncovered . . . 23
  - 0.4.1 Attention has an arrow, and default statistics of a single operator cannot see it . . . 23
  - 0.4.2 Depth is a distribution . . . 24
  - 0.4.3 Averaging can amplify, and settling is not solving . . . 25
  - 0.4.4 Length laws: sums, maxima, leaks and capacity . . . 26
  - 0.4.5 Protections the transformer already has . . . 27
  - 0.4.6 How to know what you know . . . 28
- 0.5 What holds in trained networks, and what is shown in idealized settings . . . 30
- 0.6 What this changes . . . 35
  - 0.6.1 For architecture work . . . 35
  - 0.6.2 Predictions stated in advance . . . 36
  - 0.6.3 Limits . . . 37
  - 0.6.4 Components and roots . . . 38

R Reading guide **39**
R.1 Where each part of the argument is developed . . . 39
R.2 How to read the evidence . . . 42
R.3 How the chapters depend on each other . . . 42
R.4 Notation . . . 42

I The object **44**

1 An operator read as a walk **45**
1.1 The object: an averaging kernel rebuilt from its input . . . 45
1.2 The walk's own geometry . . . 46
1.3 Amplifiers are lotteries over path lengths . . . 47
1.4 Adaptive mixing: the iteration Parts II and III study . . . 49
1.5 Three walks and a toolbox to carry along . . . 50
1.6 What holds in trained networks . . . 50

2 Depth is a distribution **51**
2.1 Architectures as path laws . . . 51
2.2 The head decides retention . . . 54
2.3 Collapse as a race between the lottery and a leak . . . 58
2.4 The mixing budget sets the rate . . . 61
2.5 Causal operators, every operator, and residual stacks . . . 63
2.6 Normalization placement is a choice of budget . . . 67
2.7 Measured: graphs and pretrained language models . . . 69
2.8 What no path law can do . . . 82
2.9 What holds in trained networks . . . 83

3 Direction: preference, exchange and circulation **85**
3.1 Attention as a stationary Markov operator . . . 85
3.2 Preference, exchange and circulation . . . 86
3.3 Sinks and cycles: the score-level criterion . . . 88
3.4 Exchange smooths and circulation transports . . . 89
3.5 Parity: quadratic readouts cannot see direction . . . 91
3.6 Measured: order at small flow and size at native flow . . . 94
3.7 What holds in trained networks . . . 96

4 One theory across graphs, denoisers and context mixing **98**
4.1 Graphs and label propagation . . . 98
4.2 Iterative denoisers and test-time context mixing are the same iteration . . . 99
4.3 Where gain should sit: placement . . . 101
4.4 A strong signal need not be useful: when propagation helps frozen features . . . 103
4.5 What the graph brings: consistency amplifies class structure . . . 105
4.6 What holds in trained networks . . . 105

II What attention needs **107**

5 Short paths: the anchor **108**
5.1 The anchor under its many names . . . 108
5.2 Removal signatures . . . 109
5.3 Pre-LN criticality and the Post-LN ledger . . . 110
5.4 The tether: consensus and coalescence . . . 111
5.5 Why capacity cannot buy back the anchor . . . 112
5.6 The class boundary: depth without a residual . . . 116
5.7 What holds in trained networks . . . 116

6 A free content path: alignment and separation **118**
6.1 The separation theorem . . . 118
6.2 The commutator of the two orders . . . 122
6.3 Address and cargo . . . 122
6.4 Leaks can grow with path length; exact block separation stops them . . . 124
6.5 Leaks in trained and pretrained models . . . 126
6.6 The tying tax and its buy-back . . . 126
6.7 What holds in trained networks . . . 128

7 Gain below the anchor **129**
7.1 The certificate . . . 129
7.2 Softmax's feedback constant . . . 131
7.3 Weight decay as the anchor . . . 132
7.4 Divergence flags do not test the coupling . . . 133
7.5 Norm control manages gain without certifying it . . . 133
7.6 Gain control is uniqueness . . . 133
7.7 Mechanism and onset in trained language models . . . 134
7.8 What holds in trained networks . . . 135

8 A way to deliver nothing: abstention **136**
8.1 Sum-to-one leaves no direction for attending less . . . 136
8.2 Every null patch is a gate times a softmax . . . 137
8.3 Rollout measures survival, and sinks act as skip switches . . . 138
8.4 The null's test decides abstention . . . 140
8.5 What trained models already show . . . 142
8.6 Own-share floors . . . 147
8.7 What holds in trained networks . . . 147

9 A margin that grows with length: the routing budget **152**
9.1 The exact simultaneous budget . . . 152
9.2 The mean-field law and tempering . . . 154
9.3 Two budgets: evidence and gain . . . 157
9.4 Sums need a log N schedule; maxima need none . . . 165
9.5 Routing information is conserved . . . 167
9.6 Retrieval's budget on length and resolution . . . 171
9.7 What holds in trained networks . . . 175

10 An arrow: circulation and order **176**
10.1 Direction requires circulation . . . 176
10.2 Canonical interventions on circulation . . . 177
10.3 A census of circulation . . . 180
10.4 Learned circulation is consequential . . . 183
10.5 Circulation implements routing . . . 190
10.6 Reversible mixers composed in depth . . . 192
10.7 The price of order . . . 194
10.8 Decoders: the leak rule . . . 196
10.9 What holds in trained networks . . . 197

11 A writer: where features are made **199**
11.1 Transport cannot write . . . 199
11.2 Invertible, volume-preserving and isometric blocks . . . 200
11.3 Retrieval: additive coupling solves it, the tested orthogonal transport does not . . . 200
11.4 Classification: writing restores parity . . . 202
11.5 Reconstruction saves memory . . . 202
11.6 Against established reversible and orthogonal designs . . . 202
11.7 The MLP as the writer and the questions left open . . . 205
11.8 What holds in trained networks . . . 205

III When averaging feeds back **206**

12 Converged averages can amplify **207**
12.1 Positive averages with a signed, amplified response . . . 207
12.2 Ordinary softmax kernels settle while amplifying . . . 211
12.3 A photographic denoiser amplifies . . . 213
12.4 Converged mixing can overturn class margins . . . 214
12.5 Own and companion margins . . . 220
12.6 Anatomy of a focal failure . . . 223
12.7 When mixing helps: guidance . . . 223
12.8 Beyond cached features: prevalence . . . 223
12.9 What holds in trained networks . . . 226

13 Settled is not solved **227**
13.1 Four readings of a fixed-point computation . . . 227
13.2 Settled but under-resolved . . . 227
13.3 A settled state need not be unique . . . 229
13.4 A root is not an attractor . . . 229
13.5 Finite computation can change outer stability . . . 229
13.6 Outer loops: directional probes miss expanding modes . . . 233
13.7 What EMA moves and what it does not . . . 234
13.8 Two loops that are safe alone can explode together . . . 234
13.9 In trained systems . . . 239
13.10 A check battery for fixed-point computations . . . 239
13.11 What holds in trained networks . . . 242

IV Why the transformer sits where it does **243**

14 The lattice and the sweet spot **244**
14.1 How the standard block pays for each condition . . . 244
14.2 The lattice: a relaxed constraint obliges a protection . . . 247
14.3 The positivity trade: signed mixing . . . 247
14.4 The price list . . . 249
14.5 The sweet spot is an interpretation . . . 250
14.6 What holds in trained networks . . . 250

15 A flat neighbourhood: compensation and knobs **251**
15.1 Compensation . . . 251
15.2 Where compensation did not apply . . . 254
15.3 Consequences: architecture search, pruning and LoRA . . . 254
15.4 An explicit knob does not beat the block's own mechanism . . . 254
15.5 Signed attention is already inside the transformer . . . 257
15.6 Does compensation grow with scale? . . . 263
15.7 What holds in trained networks . . . 263

16 Corners, rediscovery and the blueprint **264**
16.1 Pricing protections . . . 264
16.2 From concave prices to corner designs . . . 265
16.3 What the recorded designs cannot establish . . . 267
16.4 Jump or drift? . . . 268
16.5 Rediscovery . . . 268
16.6 From map to blueprint . . . 270
16.7 The scaled margin has not yet been replicated . . . 272
16.8 What holds in trained networks . . . 273

V Where the design can still move **274**

17 Abstention and retrieval at long context **275**
17.1 No fixed rule keeps both sides . . . 275
17.2 A controlled pilot . . . 276
17.3 A production model . . . 283
17.4 Trained vision transformers and 125M language models . . . 285
17.5 Maxima instead of sums . . . 292
17.6 Recommendations and limits . . . 292
17.7 What holds in trained networks . . . 292

18 Bounded memory: recall costs capacity **295**
18.1 A successor class retraces the lattice . . . 295
18.2 Strictly contracting writes bound the state . . . 297
18.3 Per-head capacity limits recall, however far back . . . 298
18.4 Hybrids keep an attention reserve . . . 300
18.5 Qualify the baseline . . . 300
18.6 In-sample length robustness . . . 301
18.7 What holds in trained networks . . . 303

19 Beyond the transformer: machines and the price list **304**
19.1 Seven conditions that do not assume a transformer . . . 304
19.2 Machines pay in different currencies . . . 305
19.3 Other substrates . . . 306
19.4 Small trained tests and machines at scale . . . 307
19.5 The room inside the transformer . . . 308
19.6 What any successor must supply . . . 309
19.7 What holds in trained networks . . . 310

20 Predictions stated in advance **311**
20.1 The predictions and their tests . . . 311
20.2 The sink length law on a production model . . . 313
20.3 The corner predictions . . . 314
20.4 The register of open tests . . . 314
20.5 What holds in trained networks . . . 316

Closing **317**

21 What this changes **318**
21.1 Rules usable now . . . 318
21.2 Limits . . . 319
21.3 Open problems and what would change the account . . . 320
21.4 What the collection does not address . . . 323
21.5 The close . . . 324

A Proofs by chapter **325**
A.1 Conventions and tools . . . 325
A.2 Proofs for Chapter 1 . . . 327
A.3 Proofs for Chapter 2 . . . 328
A.4 Proofs for Chapter 3 . . . 335
A.5 Proofs for Chapter 4 . . . 339
A.6 Proofs for Chapter 5 . . . 339
A.7 Proofs for Chapter 6 . . . 340
A.8 Proofs for Chapter 7 . . . 343
A.9 Proofs for Chapter 8 . . . 347
A.10 Proofs for Chapter 9 . . . 349
A.11 Proofs for Chapter 10 . . . 353
A.12 Proofs for Chapter 11 . . . 356
A.13 Proofs for Chapter 12 . . . 356
A.14 Proofs for Chapter 13 . . . 359
A.15 Proofs for Chapter 14 . . . 361
A.16 Proofs for Chapter 15 . . . 361
A.17 Proofs for Chapter 16 . . . 361
A.18 Proofs for Chapter 17 . . . 362
A.19 Proofs for Chapter 18 . . . 362
A.20 Proofs for Chapter 19 . . . 363
A.21 Proofs for Chapter 20 . . . 363

A.22 Proofs for Chapter 21 363

B Evidence ledger **364**

C Notation **376**

D Records and packages **382**

Bibliography **385**

Chapter 0

The Transformer Abides

0.1 The question

Nine years after it was introduced (Vaswani et al., 2017), the transformer is still the default model for sequences. It has not lacked challengers, and several have beaten it on one axis: linear attention and state-space models on the cost of inference (Katharopoulos et al., 2020; Gu et al., 2022; Gu & Dao, 2024), recurrent memories on the memory cost of long contexts (Schlag et al., 2021; Yang et al., 2024c, 2025), and sparse attention on long-context compute (Yuan et al., 2025; Lu et al., 2025). None has replaced it. Practice instead kept the design and changed it in small, specific ways. Normalization moved in front of each sublayer (Wang et al., 2019; Xiong et al., 2020); queries and keys were normalized and logits capped (Henry et al., 2020; Gemma Team, 2024); sink tokens, registers, learned sink logits and output gates were added (Xiao et al., 2024; Darcet et al., 2024; OpenAI, 2025; Qiu et al., 2025); logits were rescaled with context length (Chiang & Cholak, 2022; Peng et al., 2024); hybrids kept a fixed fraction of attention layers (Lieber et al., 2024; Qwen Team, 2025; Kimi Team, 2025); the cache was compressed into a latent representation (DeepSeek-AI, 2024); and looped models re-inject their input at every step (Bansal et al., 2022; Geiping et al., 2025).

The program behind this chapter began from four questions: how did we land on transformers; why do they seem to work so well for so many things; why have they been a fixed point, or a strong attractor, in the space of architectures; and why have we not moved on. We ask them as two. *Why transformers?* asks about the design's origin and its generality. *Why are we still here?* asks about its stability, and about where it can still move. We do not claim to answer either fully. We claim a partial answer, with a stated standard of evidence, and we say which parts of it are proved, which are measured, which are interpretation and which are still pending.

What the usual answers leave open. The familiar explanations are real, and Section 0.2 presents them with their best evidence. The transformer turned sequence modelling into large matrix multiplications, the operation accelerators improved fastest; scaling laws made capability predictable and purchasable; kernels, serving systems, recipes and people co-evolved around it; attention keeps every past token and so supports exact recall; and the block trains. These explain why this lineage was found first and why it was funded. They leave four things open. Why do most published modifications barely move accuracy: why is the design's neighbourhood flat? Which of its properties are essential, and why exactly those? Why did the changes that practice adopted take the forms they did? And where can the design still move, and what must a successor supply?

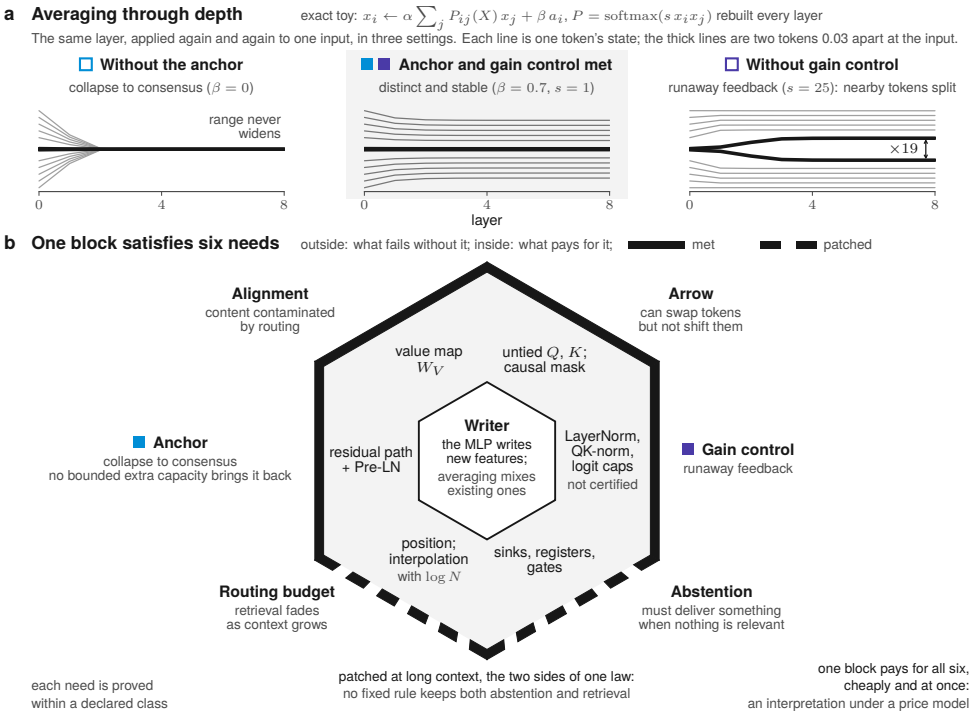


Figure 0.1: **Averaging iterated through depth can merge or split its tokens unless its needs are met (a), and the pre-normalized block pays for all six needs (b). (a)** An exact toy: the anchored iteration of Section 1.4 (the panel's formula, with $\alpha = 1 - \beta$), applied eight times to twelve one-dimensional inputs a_i. In the middle, the anchor ($\beta = 0.7$) and gain control are met (stability margin $m = \beta - 2\alpha s R^2 = 0.1 > 0$ at logit scale $s = 1$ and input radius $R = \max_i |a_i| = 1$): the iteration has one attracting state, and the two thick tokens, 0.03 apart at the input, end 0.02 apart. Each side changes one of these settings. Without the anchor ($\beta = 0$ at $s = 1$, which also makes $m < 0$), every output is a convex combination of the previous states, and the tokens collapse to consensus (rank collapse; Dong et al., 2021). Without gain control ($s = 25$ at $\beta = 0.7$), the same input has four attracting states; in the one it reaches, the two thick tokens end 0.57 apart, so a small difference in the input moves the output nineteen times as far. The margin is sufficient, not necessary: this toy keeps a single attracting state up to $s = 8$. **(b)** The six needs are the conditions of Section 0.3.2, each proved within a declared class (Table 0.2); outside each side is what fails without the need, inside is what in the block pays for it, the writer at the centre lies outside the six, and blue and violet mark the two needs drawn in (a). The dashed sides at the bottom are patched rather than met at long context, where the design can still move (Section 0.3.7): with bounded logits and a null option (to deliver nothing) that absent and present queries share, no fixed rule keeps both abstention and retrieval (Section 9.1). On the right, gain control is met by norm control, which is not a certificate (Section 7.5). On the left, under strict mixing, no extra capacity of bounded gain brings back what is lost without the anchor (Section 5.5). The toy is an illustration, not a measurement of a trained model: its anchor re-injects the input and keeps a floor, whereas at initialization the block's residual path with Pre-LN slows the loss of distinctions to a polynomial in depth and sets no floor (Section 2.6). That the block pays for all six cheaply and at once is an interpretation under a price model, not a measurement of prices.

Our answer. The collection treats a token-mixing layer as an averaging operator whose weights are rebuilt from the layer's own input, applied again and again through depth. Such a layer faces six failure modes: collapse toward a consensus, as every token drifts toward a common average; contamination of content by routing, when what decides where to look is mixed into what is carried; runaway feedback, since the weights are computed from the tokens they update; the inability to abstain, to deliver nothing when nothing is relevant; dilution as context grows, as more tokens compete for the same attention; and the inability to represent direction, to move information one way along a sequence. Each has a protection proved within a declared class of models, and a removal signature, measured at small scale: the failure that appears when that protection is taken away. For gain control, which protects against runaway feedback, the measurement is synthetic. A seventh property, writing new features, lies outside the six failure modes and their protections (Section 0.3.2, Table 0.2). We read softmax attention in a pre-normalized residual block as the one known point where all six protections are cheap at once under the prices of parallel hardware and gradient descent. That reading is an interpretation under a price model, at prices that co-evolved with the design, not a measured optimum (Section 0.3.3). Hardware sets the prices; the protections, as conditions on the block, are what is being priced.

Two mechanisms, measured at small scale, help explain why the block's neighbourhood is flat: trained networks make up for a weakened component with spare capacity elsewhere (compensation), and turning a protection into an explicit, adjustable setting (a knob) does not beat a block that already supplies it (Section 0.3.4). The anchor, the residual or input path that protects against collapse toward a consensus, is the one protection whose loss compensation provably cannot repair: under strict mixing, where repeated averaging contracts token differences toward zero, no extra capacity of bounded gain (bounded sensitivity) brings back what is lost without it, and in small-model ablations nothing else in the block made up for it. That is proved; a trained test that adds capacity is pending. Why would designs jump between discrete choices rather than drift smoothly? If design prices are concave, so that a blend of two allocations costs at least the average of their prices, a cheapest way to provide a fixed total of protection sits at a corner of the feasible set, where every protection but at most one is either off or at its cap; designs would then jump between corners rather than drift. The premise that design prices are concave is unmeasured, so this step is a hypothesis (Section 0.3.5).

Read retrospectively, many changes adopted in the mixing layer and its block since 2017 re-supply the same protections, and most of these read as responses to the failure each one fixes; the account is silent on others, such as changes to the MLP (Section 0.3.6).

The room that remains is where a protection is patched (met only by a fix that works over a limited range, such as the trained length, not by the block's structure at every length) or paid at a price that another design could avoid: abstention and retrieval at long context, governed by a two-sided log-N law in the context length (with bounded scores, no fixed rule for abstaining keeps both as the context grows); bounded memory, where a fixed-size state caps recall at a per-head capacity; softmax decisions that sum over all competitors where reading the largest score would do; and the writer of new features, for which the collection has no theory yet (Section 0.3.7). The main open problem the framework leaves is how much of each protection's price the MLP, which writes, can absorb, and at what width (Sections 11.7 and 19.6).

What the question uncovered. Answering the question turned up results that stand on their own, for readers who do not care why transformers persist (Section 0.4). Attention has an arrow: a direction carried by circulation, the net one-way flow of attention around loops of tokens. The default statistics of a single operator, such as how strongly it pulls tokens toward consensus, cannot

see this direction. Depth is a distribution over path lengths, the number of averaging steps a signal takes through the layers, and short paths decide what survives. Adaptive averaging can settle to a fixed point and still amplify changes in its input, so settling is not solving. Softmax obeys length laws that decisions on maxima escape while the competing scores stay below the decision's threshold. And the transformer already contains protections that proposals add explicitly.

What holds in trained networks. Many of the collection's results are proved for an idealized object: one operator applied over and over, a background drawn independently or replaced by its mean, keys constructed by hand. A trained network applies a different, input-dependent operator at every layer, with residual paths and MLPs between. Section 0.5 says, claim by claim, which results hold exactly for every network of their class, which hold under premises that must be checked network by network, which are shown in an idealized setting, and which are measured in trained networks, at what scale.

What this chapter does and does not do. It is the entry point to a collection of papers and the opening of the monograph that gathers them (Bruce, 2026a). Its claims are made and evidenced in the monograph's chapters, which it cites by section for every result except two remarks on graph-consistency regularization, marked as unreleased work. For three results, the weight-decay cap on the logit scale, the divergence detector's confound and the row-by-row record of the knob study, the monograph states the result and defers its full record to a later version. It re-proves nothing and reports no new experiment. Each of the collection's claims carries one of five labels (proved, measured, partial, pending, interpretation), and the usual answers this chapter reviews carry a sixth, conventional (Section R.2). The decisive tests that are still pending are named, with what each would change. It does not propose a better architecture. The account expects successors to win mainly on resource axes, such as memory, length and speed, at matched quality (Section 0.6).

Contributions.

- A fair account of the usual answers, with what the collection adds to each: support, a mechanism, a refinement, a correction, an adjacent result, or nothing (Section 0.2, Table 0.1).
- The conditions account of why the transformer persists, step by step, each step with its strongest evidence and its scope (Section 0.3).
- Where the design can still move, and what any successor must still supply (Section 0.3.7).
- A guide to what the program found beyond the question, by theme (Section 0.4, Table 0.3).
- A statement of what transfers: which results hold exactly for trained networks, which under premises, which in idealized settings, and which are measured, with the models and the scale (Section 0.5, Table 0.4).

Section R.1 introduces the monograph and suggests reading paths through it, Section R.2 sets out the evidence standards, Section 20.4 lists the tests still pending, and Appendix B lists every claim with its label and its source.

The reading guide that follows this chapter maps its argument onto the chapters (Section R.1) and says how to read the evidence labels it uses (Section R.2).

0.2 What the usual answers explain

The question of why the transformer has lasted is old, and the field has good answers to it. They fall into three families: why this lineage won, which properties of the design matter, and how the field behaves around an incumbent. We give each its best case, because the account in Section 0.3 is meant to complement them, not to replace them. Two recent papers take stock of the transformer from other sides. One considers the challengers' perspective: *The End of Transformers? On Challenging Attention and the Rise of Sub-Quadratic Architectures* (Fichtl et al., 2025). The other considers its failure modes in safety-critical use: *In Transformer We Trust? A Perspective on Transformer Architecture Failure Modes* (Mondal & Jagtap, 2026).

0.2.1 Why this lineage won: hardware, scale and history

The transformer turned sequence modelling into large, dense matrix multiplications, the one operation that accelerators were improving fastest, and ideas tend to win when they suit the machines of their day (Hooker, 2021). Loss was then found to fall as a smooth power law in model size, data and compute (Kaplan et al., 2020; Hoffmann et al., 2022). This made capability predictable and purchasable: buying scale became a safer bet than inventing architectures, and research risk became capital expenditure. Kernels, serving systems, training recipes and people's intuitions co-evolved around the design (e.g., Dao et al., 2022), so a transformer's compute became cheaper and less risky than a challenger's. Proving an alternative is expensive in its own right: wins at small scale often fail to hold at large scale (Tay et al., 2023), and a frontier run is too costly to gamble on an unproven design.

History did its part. Attention entered machine translation to align a decoder with its source (Bahdanau et al., 2015). The 2017 contribution, a design found where translation data and accelerators met, was largely a subtraction: attention without recurrence or convolution (Vaswani et al., 2017). Next-token prediction then let modality after modality be cast as one problem. Benchmarks grew up around what the incumbent made possible, which can make it look better suited to the tasks than it is. And in-context learning, picking up a task from examples in the prompt (Olsson et al., 2022), turned one model into a platform, whose use funded the next round of scale.

0.2.2 Why this design: the properties that matter

Turning to the design's own properties, one common answer is that attention has a weak inductive bias, assuming almost nothing about which inputs interact. On this view, general methods that exploit computation eventually beat methods that build in what we know (Sutton, 2019), as vision transformers showed once data was plentiful (Dosovitskiy et al., 2021). A second is the training signal: with a causal mask, one parallel pass over a sequence yields a prediction loss at every position (Vaswani et al., 2017). A third is memory. Attention keeps every past token and looks up what it needs when it reads, while fixed-size memories provably cannot recall arbitrarily long inputs (Arora et al., 2024a,b; Jelassi et al., 2024), which is why efficient hybrids keep some attention layers. A fourth is the block's organization: each layer reads from a shared residual stream and writes an edit back (Elhage et al., 2021); attention moves information between tokens, and the MLP transforms each token and can be read as a key-value memory (Geva et al., 2021).

Then come trainability and the form of the operations. Much of the 2017 design is conditioning, through scaled dot products, normalization and residual paths. Later refinements of that conditioning, such as moving the normalization before each sublayer, were about optimization (Xiong et al., 2020).

On this view architectures are selected by their loss landscapes as much as by what they can express. Two further answers point to the form of attention itself: even the input-dependent part of attention is produced and consumed by matrix multiplications, and softmax can concentrate almost all of its weight on one position, which exact copying and retrieval need (Jelassi et al., 2024). Finally, the architecture is not the algorithm: much of what would otherwise need a redesign can be learned inside the weights (e.g., Olsson et al., 2022), so the block diagram stays fixed while the computation changes.

0.2.3 Why it stays: how the field behaves

The transformer absorbs its challengers as components. Mixture-of-experts layers, state-space layers and sparse and linear attention ended up inside transformer stacks (e.g., Lieber et al., 2024; OpenAI, 2025; Qwen Team, 2025), and gating moved into attention itself (Qiu et al., 2025). Lineages built to replace it redesigned themselves for parallel training and converged towards attention (e.g., Gu & Dao, 2024), as the duality between state-space models and linear attention makes explicit (Dao & Gu, 2024). Weaknesses are worked around above the architecture, with chain of thought, retrieval and tools (e.g., Merrill & Sabharwal, 2024).

The design space around the transformer is flat: when dozens of published modifications were re-implemented in one codebase, most did not help (Narang et al., 2021). Scale sorts architectures, so far in the transformer's favour: alternatives that match it at small scale tend to fall behind as compute grows (Tay et al., 2023). And what has lasted may be a template rather than a block, since almost every component has changed since 2017 (e.g., Labonne, 2026; Xie et al., 2025).

Table 0.1 lists the twenty-one explanations above, what the collection says about each (its bearing), and where this chapter argues for that assessment.

0.2.4 What they leave open

Between them the usual answers explain why this lineage was found first, why it was funded, and why recall keeps attention in efficient hybrids. They leave four questions open.

- **Why is the neighbourhood flat?** That most modifications do not transfer, and that small-scale wins often fade with scale, is observed (Narang et al., 2021; Tay et al., 2023); it is not explained.
- **Which properties are essential, and why exactly those?** Calling the residual stream "a bus" is an intuition without a criterion. Recall explains why hybrids keep some attention, and trainability why the residual stream stays. But nothing in the usual answers says why exactly the residual stream and attention for token mixing were never swapped out, while the MLP kept its role but changed its form.
- **Why did the adopted changes take the forms they did?** Pre-normalization, normalized queries and keys, capped logits, sinks, registers, output gates, log-N scaling with context length and hybrids at fixed attention fractions are small and specific. No usual answer addresses why these forms were adopted, and not others.
- **Where can the design still move, and what must a successor supply?** The usual answers guess from costs: inference, the memory rent of a long context, the data wall. They do not say which parts of the design can move, or what a successor must still supply.

The collection is silent on much that the usual answers cover: scaling laws and their economics, data, capital and hardware prices, in-context learning beyond what single heads and short compositions do, chain of thought, and what the MLP computes. The conditions account in Section 0.3 is

Table 0.1: The usual answers and what the collection adds: for each explanation, whether the collection supports, explains, refines, corrects, is adjacent to or is silent on it, and where in this chapter the bearing is argued.

The usual answer	Bearing	What the collection adds	See
Why this lineage won			
It fit the hardware: large dense matrix multiplications	refines	Hardware sets the prices; the collection says what is being priced.	Section 0.3.3
Scaling laws made capability predictable and purchasable	silent	The collection has no scaling-law results.	—
The ecosystem co-evolved around it	refines	Not disputed; each alternative that relaxes a constraint of the block must also re-supply the protection it gives up.	Section 0.3.3
Proving an alternative at scale costs too much	refines	Compensation (spare capacity elsewhere buying back a deficit) predicts that deficits cost less at scale (pending).	Section 0.3.4
History: the right objective, data and hardware in one place	refines	The path was historical; the endpoint looks constrained, as a reading.	Section 0.3.6
We measure what transformers are good at	adjacent	Diagnostics invariant under a natural involution (such as time reversal or swapping two classes) cannot see direction or contrast.	Section 0.4.1
In-context learning turned one model into a platform	silent	Outside the collection.	—
Why this design			
Weak assumptions plus abundant data win	corrects	The premise is wrong in a precise sense: attention's bias is averaging, with failure modes the rest of the block exists to prevent.	Sections 0.3.1 and 0.3.2
Training is parallel, and every position gives a signal	silent	Beyond this, the collection notes only that contracting recurrences also parallelize and that wall-time claims need verified kernels.	—
Exact, content-addressable memory makes recall possible	supports	Supported and sharpened: a linear memory recalls N arbitrary values exactly only if $N \leq d_k$, its key dimension; recall rests on one condition of six, the routing budget.	Sections 0.3.2, 0.3.7 and 0.4.4
The residual stream is a bus; blocks communicate, then compute	explains	The residual is the anchor, the one protection that capacity cannot buy back at bounded gain (bounded sensitivity) under strict mixing; the MLP is the writer.	Sections 0.3.2 and 0.3.4
It trains: conditioning is half the design	explains	"It trains" becomes two conditions: the anchor (Pre-LN, critical at initialization) and gain control.	Section 0.3.2
Data-dependent routing built from matrix multiplications	refines	Part of the price list; other hardware would price other machines differently.	Section 0.3.3
Softmax can sharpen into a pointer	refines	Refined: that a winning score's margin must grow like $\log N$ over N keys is known (Nakanishi, 2025; Veličković et al., 2025); the collection adds its abstention side and the joint budget; with independent scores whose likelihood ratio rises with the score, no router beats the argmax on the same scores; sharpness trades against abstention.	Sections 0.3.2 and 0.4.4
The architecture is not the algorithm	explains	This is compensation, measured at small scale, with an exception this answer does not anticipate: the anchor.	Section 0.3.4
Why it stays			
It absorbs its challengers as components	explains	Most of what gets absorbed into the mixing layer re-supplies a named protection (a reading); mixtures of experts, a change to the MLP, lie outside the lattice of mixing operators.	Section 0.3.6
Challengers converge towards attention	explains	Within the declared model classes, successors must supply the same conditions and retrace the lattice of mixing operators.	Section 0.3.7
Weaknesses get worked around above the architecture	adjacent	State tracking beyond counting needs non-commuting transitions (Grazzi et al., 2025); otherwise the collection is silent.	—
The design space around it is flat	explains	Explained and measured at small scale: compensation and tests of explicit knobs for protections the block already has.	Section 0.3.4
Scale sorts architectures, so far in the transformer's favour	refines	A mechanism, as a hypothesis: compensation grows with capacity (pending).	Section 0.3.4
The transformer endures as a template for learned computation	refines	What is fixed is the set of conditions and the cheapest current way to pay them (a reading).	Section 0.3.3

complementary to the usual answers: hardware sets the prices of operations, and the conditions on the block are what is being priced.

0.2.5 Attention's strong bias is averaging

One premise in the usual answers, that attention's inductive bias is weak, is wrong in a precise sense. Attention's bias about *which* tokens interact is weak; its bias about *how* they interact is strong. A softmax head averages. Its outputs are convex combinations of its values, so they stay inside the values' convex hull, the smallest convex region that contains every value (Section 14.3). Applied again and again through depth with one fixed operator, averaging pulls every token toward a consensus unless a share of the signal keeps taking short paths, whose number of averaging steps stays bounded (Sections 2.2 and 2.3). Because each row of attention weights sums to one, a head cannot abstain (choose to contribute nothing) independently of its values (Section 8.1). Read as a random walk that steps from token to token with the head's attention weights as probabilities, a head without circulation (net traffic around token cycles at equilibrium) can swap pairs of tokens but cannot shift tokens around a cycle. Along an open sequence, such a head can route forward reliably only by piling stationary mass (the equilibrium probability assigned to each token) exponentially toward one end (Section 10.1). Much of the transformer block exists to counter the failure modes of averaging, which is where Section 0.3 begins.

Each usual answer's bearing is developed where its claim lands, in Chapters 1–3, 5, 7, 9, 11, 14–16, 18 and 19; the three answers that lie outside the collection are acknowledged in Section 21.4. The averaging premise of Section 0.2.5 is where Chapter 1 begins.

0.3 A sweet spot of conditions

This section makes the collection's case for why the transformer persists, one step at a time. Each step states its claim, its strongest evidence with its evidence label, and its scope. The steps build on one another: the object, averaging rebuilt from its input and iterated (Section 0.3.1), fixes the conditions (Section 0.3.2), and the standard block is read, under a price model, as the sweet spot where all the conditions are cheap together (Section 0.3.3). The remaining steps help explain why the design has stayed there and show where it can still move.

0.3.1 The object: averaging, rebuilt from its input and iterated

A token-mixing layer computes, at every position, a weighted average of value vectors. Its weights form a row-stochastic kernel $P(X)$ built from the current representations X, so the layer is an averaging operator rebuilt from its own input, and the block applies it again and again through depth, inside a residual path. Two proved facts turn this description into a tool. First, the polynomial filters that remain averaging operators for every row-stochastic operator (combinations of its powers, as through depth, restarts or a residual path) are exactly path lotteries, random walks of random length, $f(P) = \mathbb{E}[P^K]$ (Section 1.3). Thus what a design keeps through depth can be computed from its law of path lengths. Second, the outputs of nonnegative mixing with unit row sums stay in the convex hull of the values (Sections 11.1 and 14.3). This is why a head must deliver a mix of its values even when nothing is relevant, and why a frozen average cannot amplify: with its weights held fixed, no output moves further than the largest change among the values (Section 1.4).

The same object appears in graph propagation, label propagation, adaptive denoising and test-time mixing, and the same laws govern it there. One path-lottery law covers oversmoothing in graph networks (Oono & Suzuki, 2020), the decay of label propagation and the rank collapse of pure

attention repeated with one fixed operator (Dong et al., 2021), each known on its own (Sections 2.3, 2.5 and 4.1). On the propagation filters realized on a citation graph, the collapse-ratio identity (which gives the ratio of consensus gain to content gain from the path-length law) holds to machine precision. By contrast, a four-layer perceptron, a negative control because it does not mix across samples, shows no collapse across samples. This control is weak: it has one seed, and the deeper perceptrons tested did not converge, their final-layer features concentrating (Section 4.1).

When the mixing weights are rebuilt from the state, a second law governs stability. Consider a loop that retains a fixed share of its initial input (the anchor) at every step and averages the rest with a stochastic operator rebuilt from the current state. One anchored-contraction certificate shows that this loop contracts to a unique fixed point when its sensitivity bound holds, meaning that the rebuilt weights change slowly enough with the state relative to the anchor share. Such loops include regularization-by-denoising with a nonlocal-means denoiser and an identity observation operator. The certificate's proof is checked in Lean. p-Laplacian graph networks, the graph-diffusion model nonlinear GRAND++ and deep equilibrium transformers share this feedback between state-rebuilt weights and a retained input, but need bounds on their additional terms (Sections 4.2 and 7.1). The certificate extends to restart shares that depend on the state, the first of these additional terms: the cost in the contraction margin is how fast the share moves with the state times how far the mixed state can sit from its anchor (Section 7.1). Beyond convergence, laws of the same operator govern propagation over frozen features, test-time context mixing and, in unreleased work, graph-consistency regularization (Sections 4.2, 4.4, 12.4 and 12.5).

These laws do not depend on the domain, which is consistent with the design travelling across domains, though it does not say why attention, rather than another averaging design, was the one that travelled. The claim is not that averaging explains every domain's success, nor in-context learning; heads also compose, and the MLP writes new features. The laws above are stated for one operator, or one adaptive map, applied repeatedly, while a trained stack applies a different operator at each layer; Section 0.5 says which statements carry over.

0.3.2 What fails without each of the six conditions

An averaging layer that is rebuilt from its input and iterated faces six failure modes. Each has a protection that is proved within a declared class and a signature that has been measured when it is removed (Table 0.2). These are conditions on a class of models, not six mechanisms that every successful model must implement: they say what goes wrong, and at what price, when a model in the class lacks the property (Section 14.1).

The anchor: keep short paths. Averaged again and again, every token drifts toward a common average. A residual stack is a sum over routes: at each layer a token's signal is either averaged or passed on, so depth acts as a distribution over how many averaging steps a signal takes (a path law), and a pattern that one averaging step shrinks by a factor μ keeps $\mathbb{E}[\mu^K]$ of itself after a random number K of steps (Section 1.3). For normalized, nonnegative path laws, a non-consensus mode (a pattern that distinguishes tokens) with eigenvalue $0 < \mu < 1$ keeps at least a fixed positive fraction of itself at every depth if and only if a fixed fraction of the paths stays within a bounded number of steps. In this class, when the per-layer gains on that mode stay bounded away from zero, the stack's gain on it is $\exp(-\Theta(S_L))$ in the mixing budget S_L, the expected number of layers that mix (Sections 2.2 to 2.4). A residual connection supplies short-path mass. It keeps a fixed share only when the mixing budget stays bounded, for example when each branch shrinks with total depth. Pre-LN sits on the critical line between a floor and exponential decay at initialization, losing modes polynomially with depth (Section 0.3.3). Removing the identity path sends a minimal block to exact

chance at depth 12 and beyond, though a constant output does not by itself separate collapse from a failure to train (Section 5.2). Post-LN, whose mixing budget grows linearly with depth, has a trainability cliff from depth 18 on CIFAR-10, bimodal across seeds. A weight-tied block trained at four loops keeps 0.904–0.959 accuracy at 96 loops when it re-injects its input, against 0.023–0.032 when it does not (Section 5.2).

Alignment: a free content path. Deciding which tokens to read and deciding what to carry from them are different jobs. An independent value map is sufficient to keep them apart. It is also necessary on an identified block model when tied values inherit the keys' nuisance, as exact tying does, because values that factor through the key map carry nothing the keys lack (Section 6.1). Neither order of amplification and projection is safe on its own: amplifying before a projection exposes the part of the signal that escaped it to whatever the projection reads, and amplifying after a projection leaves in the output a component the projection would have removed. A separate value map closes the first route, for nuisance carried in the features, but not the second (Section 6.2). In trained one-hop models on a synthetic routing-and-content task with value noise, a value map tied to the keys fails at head dimension d_b, the dimension of the content, with mean retrieval 0.29 against 0.97 for separate maps (Section 6.1). Unlike the anchor, this deficit can be bought back (Section 0.3.4).

Gain control: hold the feedback below the anchor. Attention's weights are computed from its own outputs, so it can reinforce itself. An anchored adaptive iteration applies one weight builder (the map from current states to attention weights) repeatedly, as in a weight-tied loop, with anchor share β and mixing share $\alpha = 1 - \beta$. It contracts when the anchor outweighs the feedback, $m = \beta - \alpha\kappa_f > 0$, where κ_f is the feedback's loop gain. The threshold is sharp over the class, and the proof is checked in Lean. For tied dot-product softmax, the feedback bound is $\kappa_f \le 2sR^2$ (projected scores multiply it by the projections' norms), where s is the logit scale and R the radius of the representations; the bound needs that radius, since dot-product attention is not globally Lipschitz (Kim et al., 2021) (Section 7.1). Past the margin a smooth builder can have two attracting fixed points, and at the largest response of a synthetic search the equilibrium's input response (its sensitivity to input changes, in the Euclidean norm) rises from 1.77 with frozen weights to 18.9 when the weights adapt (Sections 7.1 and 12.2); a frozen average can exceed one in that norm, though no output moves further than the largest change among the values. Normalization bounds the radius but is not by itself a certificate. The per-layer bound on κ_f holds in any stack; contraction is a statement about the iterated map (Section 0.5).

Abstention: a way to deliver nothing. Softmax weights add up to one, so a head must deliver a weighted mix of its values even when nothing is relevant (zero only when they cancel): its output is a convex combination of them. A uniform logit shift, the change that would mean "attend less", leaves these weights unchanged: it lies in the kernel of the softmax Jacobian. Every zero-value patch that lets a head abstain (a sink token or a register whose value is zero, an off-by-one null or a learned sink logit) factors as a gate times an ordinary softmax, and so does an output gate. The gate controls how much the head delivers, while the softmax selects the mixture. Sigmoid attention lies outside that class (Section 8.2). Standard causal models have no explicit way to abstain, but the causal mask makes the first token absorbing, and they route heavily to it (Xiao et al., 2024; Gu et al., 2025; Barbero et al., 2025a): in seven public causal language models at 64 tokens, the minimum row mass sent to the first token has median 0.416 over 2,700 captured operators, heads and layer averages, each averaged over 16 inputs, and a per-input census of GPT-2 small gives a

per-head median of 0.080 (Section 2.7). Letting rows abstain changes the sinks. Sigmoid attention in a small vision transformer, with mean row mass near one third, cuts the share of layers with a sink from 0.53 to 0.22 at depth 12 (lower in two of three seeds) and from 0.65 to 0.07 at depth 24 (three seeds each), at a cost of 2.5 and 1.7 points of test accuracy, paid in every seed, and an output gate more than halves the mean attention paid to low-value content tokens in a small causal language model, lowering it in each of three seeds (Section 8.5). The output gate's readout carries a caveat: because the first token absorbs a causal walk, the stationary-mass test of the sink detector is vacuous on causal operators, so this readout rests on value norms.

The routing budget: a margin that grows like $\log N$. The more tokens compete, the larger the score advantage the right one needs. A single key competing with N others at a common logit needs a margin of $\log(N(1-\eta)/\eta)$ to keep a share $1-\eta$, so with a bounded logit scale retrieval fades as the context grows, as is known for retrieval alone (Nakanishi, 2025; Veličković et al., 2025). With a null that absent and present queries share (queries without and with a matching key), the largest number of background keys that keeps abstention a^* and retrieval w^* at once is the exact budget $N^* = e^{2\tau}(1-a^*)(1-w^*)/w^*$ for content logits bounded by τ (Section 9.1). On a synthetic retrieval task trained at three lengths, the context length at which accuracy drops (the accuracy onset) grows as $e^{2.14g}$ in the trained logit gap g, a descriptive fit over fifteen models whose onsets are fixed by their three training lengths. That is faster than the weight bound's e^g, because a learned decoder can read a match whose weight has already fallen (Section 9.2). Along the image-resolution axis, where upsampled images with interpolated positions increase the patch count, a training-free $\log N$ re-scoring recovers 0.65 and 0.83 points in vision transformers trained from scratch at one resolution, 3–5% of the accuracy lost; a fitted constant temperature matched that unfitted rule and beat a $\log N$ rule with a fitted exponent (Section 9.6). In ViT-S models trained on ImageNet-100, tempering every score in the same way raised accuracy at sixteen times the training token count by at least 3.7 points in every seed of the sink rules, as a prespecified test confirmed, and tempering plain softmax gained as much (Section 17.4). No result says that capacity elsewhere supplies this margin. Inside their training length, trained causal models of 0.6 million and 125 million parameters meet it through position. With rotary positions the log of the background sum (the denominator sum over non-matching keys) grows far more slowly than $\log N$, because the positional distance kernel holds distant scores down. Beyond that length this lasts while the positional treatment keeps positions in the trained range (Section 9.2); otherwise the logit scale or a $\log N$ scaling must supply the margin. When a head is no wider than the content it must recover through its keys, the budget is also dimensional (Section 6.1).

The arrow: a circulating component in each head. Many tasks need information to move one way along a sequence, as when a head reads the token that follows a match. Viewed as a Markov chain on the tokens, a head that moves information forward reliably along a sequence needs a circulating component (net probability flow around cycles), or must pile its stationary mass exponentially toward one end. For a head without circulation (a reversible chain with a positive stationary law), the odds of moving to a successor rather than a predecessor have geometric mean exactly one around every cycle of tokens whose successor and predecessor probabilities are positive: it can swap tokens but not shift them around a cycle (Section 10.1). This is Kolmogorov's criterion for a reversible chain (Kolmogoroff, 1936; Kelly, 1979), read as a statement about what a head can do. Tying queries to keys costs more than circulation. Tied scores form a positive-semidefinite Gram matrix, which on geometric average around a cycle cannot prefer a neighbour to the reading token itself. On a synthetic cyclic-retrieval task with small models, tied models reach 60% and

76% successor retrieval at depths 2 and 4, against 98% for circulation-free heads with a symmetric indefinite score form, which is not subject to that Gram restriction, and 100% for standard ones. Under a causal mask, which gives look-back relations their direction, the Gram form still copies less surely than untied scores at varying offsets in a small synthetic test, with cross-entropy 3.4 times higher on average over five seeds and accuracy 0.8 points lower, while a bidirectional lookup stays at parity (Section 10.1). Reversing the first layer's learned circulation in trained retrieval models turns successor retrieval into predecessor retrieval (Section 10.5). The statement is per head: compositions of reversible heads can acquire direction, and decoders take theirs from the causal mask.

The writer lies outside the inventory. Moving information around is not enough. Averaging cannot create new features: a mixing step with row-stochastic weights, however they are chosen, keeps every output in the convex hull of its input rows and cannot raise their rank, while a shared pointwise nonlinearity can do both (Section 11.1). Replacing averaging with orthogonal token transport hits a similar limit through an invariant: mixing the tokens of a group by one shared orthogonal matrix preserves the group's channel-by-channel inner products (its channel Gram matrix), so a readout that depends only on that matrix cannot change, though individual tokens can. A transport-only mixer of this kind reaches a seed-mean dictionary-retrieval R^2 of 0.18–0.26 at the training size of 16 associations (below 0.03 at 64), against at least 0.99996 for standard and additive reversible blocks at 8–64 associations, and loses 10 points on CIFAR-100, with arms that are not parameter-matched; adding a nonlinear feature-writing channel to the transport mixer restores classification but not retrieval (Sections 11.3 and 11.4). In a transformer the MLP and W_O write, and the collection has no theory of what the MLP should compute (Section 11.7). Stated for routing machines in general, writing becomes a seventh condition (Bruce, 2026e) (Section 19.1).

Scope. Each row of Table 0.2 is a theorem within its class plus a measured removal signature, and this is the strongest step of the case. Its limits are specific. The classes are particular (the anchor, for example, is stated for normalized nonnegative path laws of a fixed mode), the removal signatures were measured at small scale, mostly below 25 million parameters, and the gain-control signature is synthetic. A row would be falsified by a trained model in its class that lacks the protection and shows no signature; for the anchor, a registered test of exactly this is pending. Compensation (Section 0.3.4) does not excuse a missing signature by default. It is itself a measured claim with a named currency, such as MLP width, head width or head count: a model that lacks a protection and shows no signature counts against its row unless removing that currency brings the signature back. That test exists so far for the value map; for the anchor, compensation at bounded gain is ruled out under strict mixing, and the trained test is the one pending above.

0.3.3 The sweet spot

Each part of the standard block pays for one condition (Figure 0.2). The residual stream, with normalization in front of each sublayer, is the anchor. It sits on the critical line at initialization, where a logarithmic mixing budget gives polynomial rather than exponential decay: in the mean-field energy model, a pre-normalized stack retains $L^{-s(1-\mu^2)}$ of a mode through L layers, where μ is the mode's eigenvalue and s, here, the attention share rather than the logit scale. At width 512 (three seeds), random pre-normalized transformers with bidirectional attention, whose attention is nearly uniform at initialization, match the predicted exponent $-s$ with a largest deviation of 0.041, and, at width 512, linear token-mixing stacks match its dependence on μ within 0.05. Post-LN's mixing budget, by contrast, grows linearly with depth and drives exponential decay (Section 2.6). For

Table 0.2: Six conditions and the writer: what each guarantees, what fails without it (record and signature), how the pre-normalized softmax block pays for it, whether capacity elsewhere can compensate, and the chapters of the monograph that develop it (Bruce, 2026a).

Condition	What it guarantees	What fails without it	In the block	Compensated or patched?	Chapter
Anchor	A mode keeps at least a fixed fraction of itself at every depth iff a fixed fraction of paths stays short; with per-layer gains on that mode bounded away from zero, the stack's gain on it is $\exp(-\Theta(S_L))$ in the mixing budget	Identity path removed: exact chance at depth ≥ 12, which does not isolate collapse. Post-LN cliff at depth ≥ 18. Looped block trained at 4 loops, at 96 loops: 0.023–0.032 without re-injection, 0.904–0.959 with it	Residual stream with Pre-LN, critical at initialization	No, at bounded gain under strict mixing	2 and 5
Alignment	An independent value map is sufficient, and necessary on an identified block model when tied values inherit the keys' nuisance	Tied value map (with value noise): retrieval 0.29 vs 0.97 at head dimension d_b	The value map W_V	Yes: width (0.96 at $2d_b$) and MLP capacity	6
Gain control	Contraction when $m = \beta - \alpha\kappa_f > 0$, with $\kappa_f \leq 2sR^2$ for tied dot-product softmax; sharp over the class	Past the margin, two attracting fixed points (synthetic); at a synthetic search's largest response, Euclidean response 1.77 frozen vs 18.9 adaptive	LayerNorm, later QK-norm and logit caps: norm control, not a certificate	Partly; norm control is not a certificate	7 and 12
Abstention	Rows summing to one cannot abstain independently of the values; every zero-value patch for abstention and every output gate factors as a gate times a softmax	Sigmoid attention (row mass 0.34 and 0.31 at depths 12 and 24) cuts sink incidence from 0.53 to 0.22 and from 0.65 to 0.07; an output gate halves content-sink mass (value-norm readout; causal caveat)	Sinks, registers and gates, patched inside unit row sums	Patched; at long context no fixed rule with a null shared by absent and present queries keeps both abstention and retrieval	2 and 8
Routing budget	One key among N needs a score margin growing like $\log N$; with a shared null and logits in $[-\tau, \tau]$, $N^* = e^{2\tau}(1-a^*)(1-w^*)/w^*$	Accuracy onset grows as $e^{2.14g}$ in the logit gap g; a training-free $\log N$ re-score recovers 0.65 and 0.83 pp (3–5% of the loss, matched by a fitted constant) on upsampled images	A fixed logit scale, helped by position within the trained length; beyond it, positional interpolation and log-N scaling, patched	Within the trained length, through position; no result for capacity elsewhere; dimensional at narrow heads	9 and 17
Arrow	A head without circulation can swap tokens but not shift them (per head)	Tied (Gram) scores: successor retrieval 60% and 76% at depth 2 and 4, against 98% for circulation-free indefinite scores and 100% for standard heads; under a causal mask, cross-entropy 3.4 times higher at varying offsets	Untied queries and keys; the causal mask in decoders	Yes, by composition across heads and layers	10
Writer (outside)	Averaging cannot create features: outputs stay in the inputs' hull, and rank cannot rise	Transport-only mixer: retrieval R^2 0.18–0.26 at 16 associations (below 0.03 at 64) vs ≥ 0.99996 at 8–64; CIFAR-100 −10 pp (parameters not matched)	The MLP	A write channel restores classification, not retrieval	11

alignment, the value map W_V is the free content path. LayerNorm, and later QK-norm and logit caps, control gain, by norm control rather than a certificate. Sinks, registers and output gates patch abstention inside unit row sums. A fixed logit scale supplies the routing budget within the trained length, helped in trained causal models by rotary positions that hold down background growth. Beyond the trained length, positional interpolation patches the budget. Log-N scaling, added later, helps on interpolated positions unless the interpolation rule already applies a log-length temperature (Section 0.6.1). Untied queries and keys, and in decoders the causal mask, supply the arrow. The MLP writes. The inventory was assembled by asking what each part of the block protects, so this match holds by construction; what does not follow from the construction is each row's necessity within its class (Section 0.3.2) and the predictions about other designs (Section 0.6.2).

In small trained models, tuned standard components already carry the protections credited to them and capabilities that dedicated knobs are added to supply. In graph transformers, a pre-normalized stack's last-layer anchor coefficient, an input–output regression coefficient, is 0.998 on average, against 0.934 for Post-LN (165 cells each: ten datasets, depths 2 to 16 with incomplete coverage, five split seeds), and a tuned Switch load-balancing loss already balances its experts (three seeds) (Sections 5.3 and 15.4). Multi-head attention with signed output maps already subtracts across heads: in a single small block trained for 500 steps, difference retrieval (recovering the difference of two stored values) reaches a mean R^2 of 0.46 with one head and 0.998 with four (25 matched runs) (Section 15.5).

We read this as a sweet spot: under the prices of parallel hardware and gradient descent, the pre-normalized softmax block is the one known point where all six conditions are cheap together. This is a judgment, not a computation. The price model's formal parts price spectral costs (how much collapse toward consensus each species of operator must pay to suppress an unwanted mode (Section 14.4)) and budget allocations across protections under assumed concave prices (Sections 16.1 and 16.2). Cheapness on hardware is judged, and the alternatives' prices were not measured on a common budget of hardware and training (Section 14.5). Nor are the prices independent of the design: hardware, kernels and recipes co-evolved with the transformer, so part of its cheapness is incumbency. The transformer is the cheapest point under one price list, not the only design that meets the conditions: a catalog of 43 families of routing machines shows how others pay for them (Sections 14.4 and 19.2). This is also where the account meets the usual answers of Section 0.2: hardware sets the prices; the conditions are what is being priced.

Underneath the price list sits a family of lower bounds (Section 9.5). Whatever the machine, naming one of N stored items costs about $\ln N$ nats, by Fano's inequality (Cover & Thomas, 2006), and keeping a context of unstructured items for arbitrary questions asked later costs memory in proportion to the number of items. Machines pay in different currencies: key width (dimension), sharpness of the read (logit range), state or order. Each bound holds within its own class of machines and prices one currency there; the bounds set no exchange rate between currencies. Together they say that no fixed allotment of dimension, logit range or state survives a growing context, which may be one reason efficient alternatives keep reacquiring attention-like memory.

Relaxing any constraint of the block buys a capability and obliges the design to re-supply the protection it gave up (Section 14.2). Sigmoid attention gains freedom in its row sums and must re-supply a length budget and norm control; recurrences gain a bounded state and must re-supply recall; signed mixing gains subtraction and must re-supply a cap on its norm or per-head normalization (Sections 14.2 and 14.3).

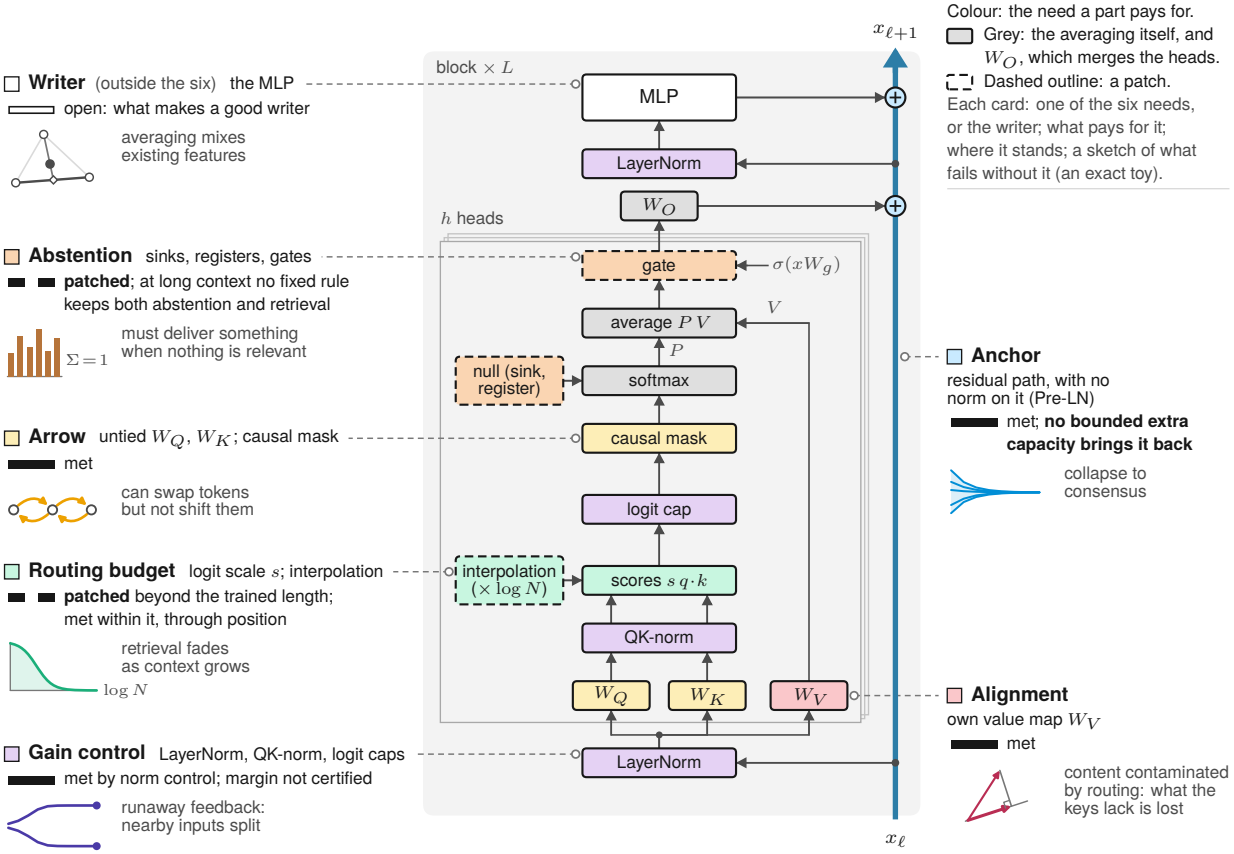


Figure 0.2: **Each of the six needs and the writer is paid for by a familiar part of the pre-normalized block; two of the needs, abstention and the routing budget, are patched rather than met at long context.** As in Figure 0.1, inside is what pays for each need and outside is what fails without it. Block ℓ of L maps the residual stream x_ℓ to $x_{\ell+1}$; s is the fixed logit scale, N the context length, and x in the gate $\sigma(xW_g)$ the normalized input; positional encoding is not drawn. The patches are positional interpolation beyond the trained length (Peng et al., 2024), with log-N scaling, drawn in parentheses, where positions are interpolated (Chiang & Cholak, 2022); a value-null option at the softmax (a key that takes weight and delivers nothing), such as a sink or a register (Xiao et al., 2024; Darcet et al., 2024); and an output gate on each head (Qiu et al., 2025). With bounded logits and a null that absent and present queries share, no fixed rule keeps both abstention and retrieval at long context (Section 9.1). Each sketch is an illustrative construction: an exact toy of the failure beside it, not a measurement. Each need, and the writer, is proved within a declared class and has a signature measured when it is removed, synthetic for gain control (Table 0.2). Under strict mixing, no extra capacity of bounded gain brings back a missing anchor (Section 5.5). Gain control is met by norm control, not a certificate (Section 7.5). The arrow is stated per head. What makes a good writer is open. One part per need is a compression: the logit scale also enters the feedback gain and W_O also writes (Sections 7.2 and 11.7), and QK-norm (Henry et al., 2020), logit caps (Gemma Team, 2024) and the patches are later additions; models differ in which they have. That the block pays for all the needs, cheaply and at once, is an interpretation under a price model.

0.3.4 Why the neighbourhood is flat

Two measured mechanisms make local changes rarely pay: when a change weakens a protection, spare capacity elsewhere can buy back the deficit (with the anchor as the exception), and when a change turns a protection into an explicit, adjustable setting (a knob), a tuned block already supplies it.

Compensation. Each protection's necessity is proved on a construction built to isolate it, and a trained network evades the theorem whenever a resource, such as width, heads or MLP capacity, is spare (Section 15.1). On a theorem's own construction for $C = 10$ classes, a content map of rank $C - 2$ sits at 90.16% against a predicted ceiling of 90.0%, while a content map of the same rank, trained jointly with its readout, reaches 99.84%; both are means over five seeds. Where a single nonnegative head cannot subtract, multi-head attention buys back subtraction across heads (Section 0.3.3) (Section 15.5). A value map tied to the keys fails at head dimension d_b (the content dimension of Section 0.3.2), and twice the width restores retrieval to 0.96. At depths 12 and 24, MLP capacity also absorbs the accuracy drop (tax) of tying the value map. In small vision transformers at depth 24 and width 192 on CIFAR-100, that tax is +1.70, +0.04 and −0.72 percentage points at MLP ratios 0, 1 and 4. The penalty shrinks as the MLP widens: the interaction between tying and capacity is negative in all four configurations at depths 12 and 24 (Holm-adjusted $p \leq 8.4 \times 10^{-6}$), while at depth 4 the tax stays positive over the measured range; the break-even MLP ratio where the tax reaches zero is 1.11, with a stratified 95% bootstrap interval of [0.93, 1.63] (Section 6.6). These results come from a reanalysis of a known outcome, over five initialization-only seeds that share one data order. The comparison is not parameter-matched: tied models have fewer parameters. This is compensation at the level of accuracy; the records do not show the mechanism. In theory, the condition for a downstream compensator to exist is settled: inputs sent to the same state by the weakened (restricted) map must have the same target value, meaning that the target is constant on the fibres of the restricted map (Section 15.1). In a trained model, an absent removal signature is put down to compensation only when removing the spare resource brings the signature back, the test stated in the scope of Section 0.3.2.

The anchor is the exception. It is the protection whose loss is proved beyond the reach of capacity elsewhere: under strict mixing, capacity of bounded gain cannot buy it back. If the layers act on a common discriminative invariant subspace with norms $q_\ell \leq 1$, any downstream recovery map with Lipschitz constant G that must separate two distinct inputs in that subspace by $c > 0$ times their distance needs $G \geq c/\prod_\ell q_\ell$ when the product is positive. Under strict mixing, where these products vanish with depth, no compensator of uniformly bounded gain recovers the lost distinctions; large gain at finite depth, an operator close to the identity or an external input can recover them (Section 5.5). Trainability already explains why the residual stream is the part of the block no one has replaced; this non-compensability adds a second reason.

An explicit knob does not beat the block's own mechanism. Twelve interventions, each registered in the repository before its runs (Section R.2), turned one protection into an explicit, certifiable knob inside a standard model: certified fast weights, an anchored mixture-of-experts router, sign-flexible filters, a sink-anchored decoder and eight more (Section 15.4). Each knob's mechanism held in several rows where its hypotheses held, several of them to numerical precision, but not in all: it was not met in one row, was inconclusive in two, and its falsifier was realized in a fourth. Certified fast weights, for example, had no divergences in 48 certified runs, where 53 of 273 uncertified runs diverged (41 of 45 aggressive and 12 of 228 tuned, the tuned ones all in language

modelling, which had no certified arm) (Section 18.2). Yet each of the ten capability tests that could be scored was null or negative; the row-by-row record is still to be released. Where a knob gained, the gain stayed below the preregistered bar or was paid for on another axis. The sink-anchored decoder, a 76.7M-parameter model trained at 512 tokens, bought robustness at length, cutting perplexity at 2,048 tokens from 88.1 to 59.8. This came at an in-distribution cost: perplexity at its training length rose from 25.80 to 26.85. These are means of two seeds, which agree in direction. The study's explanation, measured directly so far only for the anchor and a Switch router's balance (Section 0.3.3), is that a tuned baseline already supplies the protection.

Measured at small scale, these two mechanisms together offer an answer to the first open question of Section 0.2.4: a reason most published modifications barely move accuracy. If compensation grows with capacity, it would also explain why the small-scale gains of alternatives fade with scale. That is a hypothesis, and its test is pending. *Pending:* Compensation budget at scale (vision transformers on ImageNet-100 and a 125M language model): does the price of a deficit shrink with capacity? *Pending:* The anchor non-compensability sweep in trained networks.

0.3.5 Do designs jump rather than drift?

This step is a hypothesis about why designs would jump between discrete choices rather than drift smoothly. Under concave prices, some best allocation of a design budget across protections is a corner of the feasible set (a vertex), with at most one protection strength strictly between its bounds. For a fixed feasible set under those prices, the best corner changes only when the costs of competing corners cross, so the optimum jumps from one corner to another (Section 16.2). That a concave cost is minimized at a vertex is classical (Rockafellar, 1970, §32). In 36 constructed instances an interior split never beat the best corner, though knowing the optimum is a corner does not say which corner it is: under capped, anchored pricing, greedy filling chose a suboptimal corner in 5 of 24 instances.

In deployed architectures, hybrids that mix full and linear attention do sit at or next to small-denominator fractions of attention, with full attention in 12 of 48 layers of Qwen3-Next and 7 of 27 of Kimi Linear (Qwen Team, 2025; Kimi Team, 2025). But that record is descriptive, and it was compiled by those holding the hypothesis.

The vertex result's premise, concave pricing at the level of architectures, is an assumption, not a measurement. And on a single attention-fraction axis concave pricing predicts 0, 1 or a binding cap, never a fraction such as 1/4 in itself. The small-denominator fractions used in the predictions (Section 0.6.2) are a scoring convention, not a consequence of concave pricing. The deployed fractions therefore do not yet discriminate between corners and drift; a test needs measured prices, or a controlled comparison of jump against drift, which is planned (Section 16.4).

0.3.6 Adopted changes in retrospect

Read retrospectively, many of the changes adopted in the mixing layer and its block over the last decade re-supply a named protection, and most of these read as responses to the failure each one fixes (Section 16.5). Residual paths and pre-normalization re-supply the anchor and choose a mixing budget (He et al., 2016; Wang et al., 2019; Xiong et al., 2020). QK-norm, σReparam and logit soft-capping control attention's gain (Henry et al., 2020; Zhai et al., 2023; Gemma Team, 2024), and z-loss controls the output logits (Chowdhery et al., 2023). Sinks, registers, off-by-one normalization, learned sink logits and output gates supply abstention (Xiao et al., 2024; Darcet et al., 2024; Miller, 2023; OpenAI, 2025; Qiu et al., 2025). YaRN's interpolation and temperature, log-N scaling and scalable softmax pay the routing budget beyond the trained length (Chiang & Cholak, 2022; Peng et al., 2024; Nakanishi, 2025). In trained 125M language models, however, scaling every logit by

$\log N / \log N_{\text{train}}$ (the ratio of log context length to log training length) helped only when positions were interpolated without another log-length temperature (Section 0.6.1). Differential attention is signed mixing with norm control, not a certified cap on the norm (Ye et al., 2025). Contracting recurrences and hybrids at fixed attention fractions bound the state and keep a reserve for recall (Yang et al., 2025; Lieber et al., 2024; Qwen Team, 2025; Kimi Team, 2025; MiniMax, 2025). Latent key-value caches compress the content channel (DeepSeek-AI, 2024); learned sparse attention relaxes density with learned support (Yuan et al., 2025; Lu et al., 2025); and looped models with input re-injection restore the anchor along the axis of recursion (Bansal et al., 2022; Geiping et al., 2025). The chronology behind Figure 0.3, from the paper released with this one on the transformer as a constrained equilibrium of protections (Bruce, 2026d), selects thirty-three sources from thirteen such groups of proposed or adopted changes and dates them by first public release; several, among them differential attention and looped models, are proposals that practice has not generally adopted (Section 16.5).

This is a reading, not evidence. None of these changes was predicted from the conditions (Section 0.3.2) before it was adopted, so none counts in their favour. The lattice of trade-offs (Section 0.3.3) describes relaxed constraints and the protections that must then be re-supplied (Section 14.2). It is silent on much that was adopted, such as changes to the MLP, to kernels, to the sharing of keys and values across heads (multi-query and grouped-query attention; Shazeer, 2019; Ainslie et al., 2023) and to the block's topology (parallel attention and MLP branches). We searched for a counterexample: an adopted change to the mixing layer that removes a protection without re-supplying it. We found none, but the search was ours, not blind. We include the reading because it answers the third open question of Section 0.2.4 (why the adopted changes took the forms they did), which nothing else does, and because it is the most intuitive way into the account.

The same reading runs forward as a blueprint. Assembling the smallest block that meets the inventory of Section 0.3.2 gives the pre-normalized transformer, within four corrections: explicit registers, one symmetric structure map where direction is not needed, a log-N budget, and residual branches scaled by $1/L$ (Section 16.6). Since the inventory was read from the block, this shows that the account is consistent, not that it is independently confirmed. Because the symmetric structure map makes each layer reversible, in small models this minimal certified block keeps detailed balance to 10^{-12} on every layer of 50 trained encoders. The same block, completed with interleaved MLPs and pre-normalization, beat a parameter-matched ViT-Tiny after 100 epochs in 20 of 20 paired runs on CIFAR-10 and CIFAR-100, by 1.8 to 3.9 points averaged over the five seeds of each dataset and learning rate (Section 16.6). At 22 million parameters, the recorded margins over ViT-Small were 10.59 and 8.02 points of top-1 accuracy on Tiny-ImageNet and ImageNet-100, averaged over five paired seeds. These margins are not evidence of dominance: the baselines were not tuned, and the tuned replication has not run (Section 16.7). The path to the transformer was historical; its endpoint, read this way, looks constrained. *Pending:* The tuned replication of the minimal block at 22M parameters.

0.3.7 Where the design can still move

The room that remains is where a condition is patched rather than met, or paid at a price that another design could avoid (Figure 0.4): abstention at long context, bounded memory, decisions that read maxima instead of sums, and the writer.

Abstention at long context. Unit row sums forbid abstaining independently of the values, and every fix in use (a sink, a register or an output gate) sits in one cell of the design lattice. A two-sided law governs that cell. With content logits bounded by τ, no fixed, learned or length-calibrated

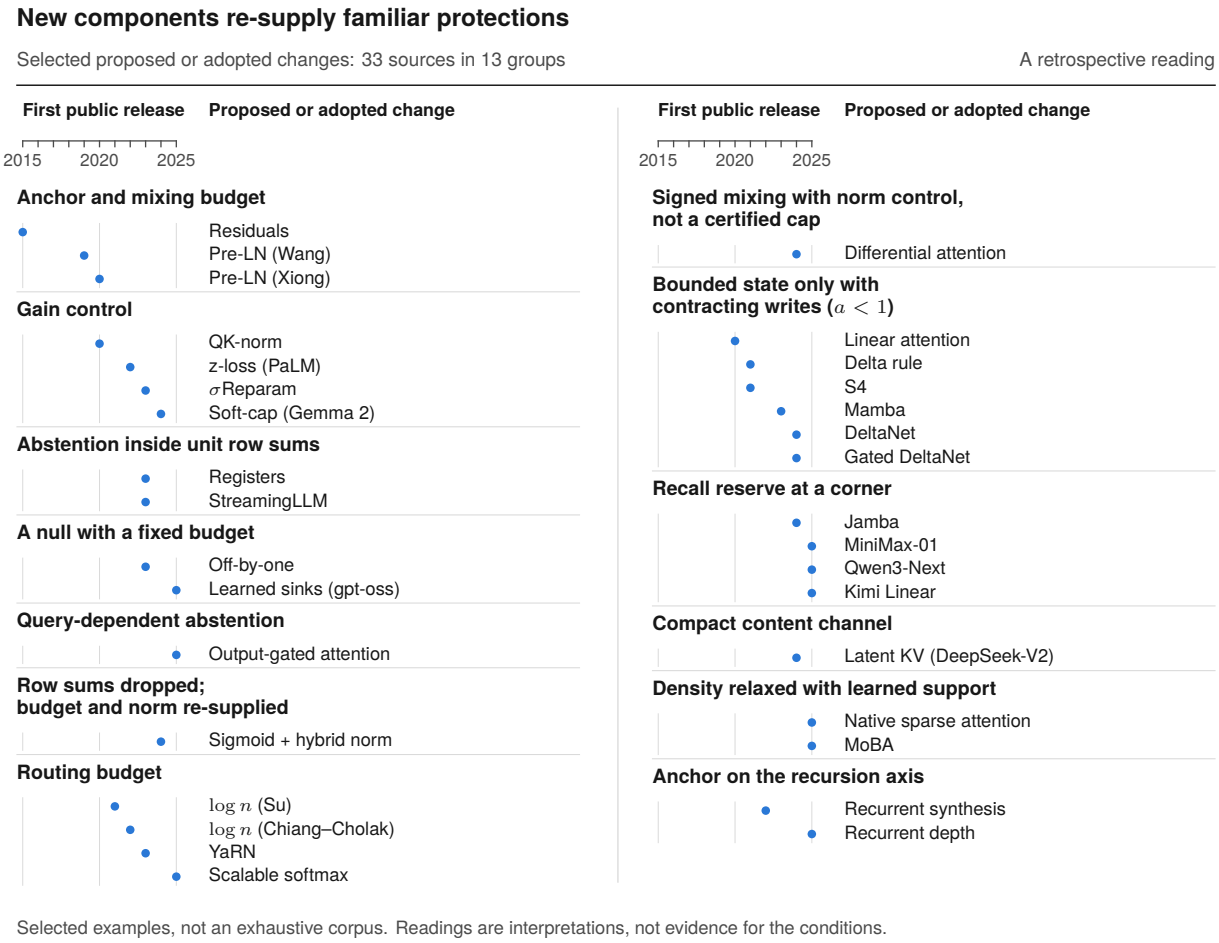


Figure 0.3: **Selected architectural changes revisit familiar protections: a retrospective reading, not evidence for the conditions.** The figure draws on a chronology of 33 separately dated sources in 13 groups (Section 16.5). Each dot marks the year of first public release (arXiv v1, blog or model card), not adoption; each row names one source, and vertical position only separates entries. The two timelines use the same scale. Group headings give the readings of the paper on the transformer as a constrained equilibrium (Bruce, 2026d); the examples are proposed or adopted, not an exhaustive corpus or a claim of global priority. Differential attention supplies norm control, not a certified cap. Bounded state requires contracting writes ($a < 1$, for a decay factor a on the state); listing a method does not certify that condition. PaLM's z-loss acts on output logits, so its gain-control placement is a block-level interpretation. Besides the works cited in Section 0.3.6, the figure plots Su (2021), Ramapuram et al. (2025), Katharopoulos et al. (2020), Schlag et al. (2021), Gu et al. (2022), Gu & Dao (2024) and Yang et al. (2024c). Exact source dates and their qualifications are kept with the figure's records.

null that absent and present queries share keeps both absent-query abstention weight a^* and present-query single-key retrieval weight w^* beyond $N^* = e^{2\tau}(1-a^*)(1-w^*)/w^*$. The logit range must grow. Tempering, which multiplies every score, the null included, by $\log N / \log N_{\text{train}}$, is one way that keeps both when the margins are strictly separated (Sections 9.1 and 9.2). The same budget decides any other schedule that scales the content scores and sets a null (Section 9.1). In a controlled pilot of trained lookup-or-abstain heads, at logit scale at most 10 and a training length of 64 keys, a fixed learned sink's abstention falls from 0.874 to 0.007 (seed means) between 64 and 65,536 keys, along the law's prediction within 0.002 per seed; length-calibrated rules lose retrieval from 2,048 keys; and the tempered rule keeps both over the tested range, empirically rather than by certificate. In a registered extension to seven other settings, every rule that sums over the keys, tempering included, lost a side with heavy-tailed keys, and at the smaller budget tempering also lost abstention with clustered keys and with four heads (Section 17.2).

Trained language models can offset a fixed sink's decay inside their training length by holding the background sum down through position. The prespecified production-model test, on gpt-oss with banded layers (whose finite window caps the background) as a built-in control, was undecided by its rule, although descriptively its slopes lean toward such compensation through position. Its books control, which compares next-token loss on natural and shuffled text, failed, and its sink log-odds slopes matched neither prespecified branch: confirmation of the law or compensation (Section 17.3). In trained language models at 125M parameters, tempering every logit by $\log N / \log N_{\text{train}}$ on top of YaRN did not keep abstention and raised long-context perplexity. YaRN already carries a log-length temperature, and none of the probe's retrieval heads had the separated margins the schedule needs. Inside their training length, these models hold abstention nearly level through position: the rotary kernel holds the background's growth down, and training widens the kernel when the null itself grows with length, as small models with and without positions confirm. Beyond the training length, the hold on text lasts while positions are interpolated back into the trained range, and fails natively (without interpolation). On the probe's absent queries it did not hold even with YaRN: the fixed sinks kept about a tenth of their null mass at eight times the training length. Every rule that gives attention a null removed the first-token sink and most massive activations, as earlier reports found for heads that can do nothing, attention biases and output gates (Bondarenko et al., 2023; Sun et al., 2024; Qiu et al., 2025). At this scale no retrieval head reached the registered identification threshold (the largest weight was 0.247, against 0.3), so these abstention results describe attention weights, not behaviour (Section 17.4).

Bounded memory. A fixed-size state stays bounded at every length when its writes contract by a common factor. A delta-rule write with decay contracts in this way, with bounded keys and values and a small enough step. Non-expansiveness alone does not ensure boundedness, and such a state can grow like $\sqrt{t}$ (Section 18.2). In a trained hybrid a decay gate acts as the anchor that supplies this contraction: without it, the log of the state's norm rises by 2.23–2.28 over a sequence of twice the training length, in the layer where it rises most, against 0.24–0.87 across the four variants that keep it (five seeds each, at the larger of two model sizes) (Section 18.1). A bounded state's recall price is a capacity, not a distance: a linear memory can recall every assignment of values to N keys exactly if and only if the keys have rank N, so exact recall of N arbitrary associations needs $N \le d_k$ per head, where d_k is the key dimension (Section 18.3). A gated delta rule recalls 32 pairs at 0.998–0.999 (five seeds); at 128 pairs, beyond the 32 it was trained on, it recalls 0.78–0.83. These records hold no attention arm for comparison, so the recall cost of bounded state is not yet established in trained models (Section 18.3). Because bounded state caps exact recall per head, a hybrid keeps an attention reserve for recall, though even one layer with full context keeps the cache

linear in the context (Section 18.4). *Pending:* The qualified hybrid ratio sweep at 125M and 350M.

Maxima instead of sums. With an exchangeable background, a decision that adds up all N competitors at a fixed logit scale must be retuned as their number grows, by a $\log N$ schedule; a decision that reads only the largest needs none, and fails only when the background maximum reaches its threshold (Sections 9.4 and 17.5). In constructed tests, the fixed softmax sink failed at the predicted onset in four of four configurations. By contrast, a key check abstains unless its best score clears a threshold; with a weak match in these tests, it outlasted $\log N$ tempering. A head built on maxima, the max-statistic head, which routes by a softmax over its top keys and abstains through a gate that compares its largest score with a learned threshold (Bruce, 2026e), should then keep abstention and retrieval at every length without a schedule, against backgrounds bounded below its threshold. In the trained 125M language models, with positions interpolated by YaRN, its null and match weights both fell by less than half a nat in log-odds from the training length to eight times it, on the seed means though not in every seed, which no other rule achieved; five rules kept the match weight alone better. Without interpolation it failed as the matching key fell far down the ranking and the background maximum grew with length, hitting the evidence wall that the theory names, where background keys outscore the match (Section 17.4).

The writer. No theory here says what a good writer is. The main open problem the framework leaves is how much of each protection's price the MLP, which writes, can absorb, and at what width (Sections 11.7 and 19.6).

What a successor must supply. Within the stated operator classes, any replacement must keep short-path mass, a free content path, gain control of the full feedback, a way to abstain whose budget survives length, a routing budget that grows like $\log N$, an arrow (from circulation, a mask or composition), and a writer (Section 19.6). We expect a successor to win on a resource axis at matched quality, not on quality at matched cost.

0.3.8 How strong this case is

The evidence for the case in Section 0.3 ranges from theorems and measurements to interpretations and pending tests, in order of weight:

1. **Proved within declared classes.** The conditions and the laws are theorems about stated classes of operators or models, except the stationary law of the tempered head in Section 0.3.7, which is checked numerically, not proved. Most are also checked numerically, and several are formalized in Lean (Section R.2). None says that every model must implement six named mechanisms.

2. **Measurements are mostly at small scale.** Removal signatures and compensation are measured in controlled training runs, mostly below 25 million parameters and none above 125 million. For gain control, the removal signature is synthetic. Evidence at production scale comes instead from measurements on public checkpoints (Sections 0.4 and 0.5): 2,700 captured operators, sixteen pretrained language models, 184,320 certified head kernels and seven production vision transformers.

3. **Idealized settings.** Many results are proved for one operator applied repeatedly, or for idealized backgrounds and constructed keys; Section 0.5 labels each.

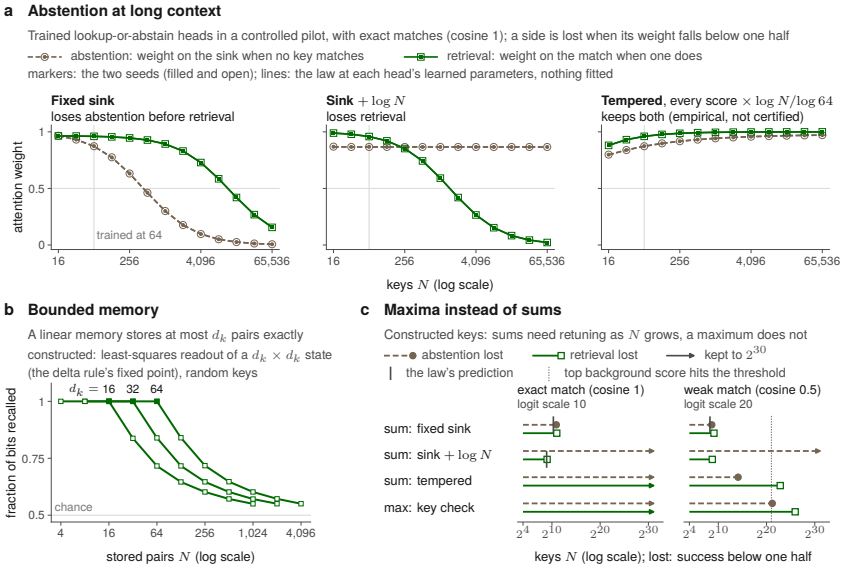


Figure 0.4: **With exact matches a fixed rule loses abstention or retrieval as N grows, while log-N tempering keeps both over the tested range (a) and a decision on the largest score needs no length schedule (c); a linear memory stores at most d_k pairs exactly (b).** (a) A controlled pilot (Section 17.2), logit scale at most 10: means over about 512 queries per point (half of the 1,024 evaluated); the two seeds within 0.002 of each other. The curves plot sink and match weights: the fixed sink follows its weight law within 0.002 in each seed, and its retrieval accuracy stays at 1 even as its match weight falls. Sink $+\log N$'s retrieval accuracy falls below 0.9 from 2,048 keys in both seeds, as does sigmoid $-\log N$'s (not drawn). The tempered head reuses the trained fixed-sink head with its scores tempered at evaluation; its law is a stationary law, checked numerically, not proved. Three of the pilot's seven rules and one of four configurations are drawn. (b) Constructed memories (Section 9.5): random unit keys and ± 1 values read by sign; 8 to 16,384 trials per point, no intervals saved. Exact recall of N arbitrary values needs $N \le d_k$ (Section 18.3); the least-squares readout (the delta rule's fixed point) reaches it (filled: $N = d_k$) and, by construction, recalls the oldest pairs as well as the newest, whereas single-pass online writes (the delta rule, linear attention) are not exact even below $N = d_k$. (c) Constructed product keys (Sections 9.3 and 9.4), from historical records, recovered but not reproduced (Bruce, 2026e, Appendix C.2); sinks calibrated at 64 keys; 6,000 queries per point (three seeds pooled, no spread kept); sizes four-fold apart, so a loss (where success first falls below one half, interpolated in $\log N$) is located to within a factor of 4; two of four configurations drawn. The key check abstains unless its best score clears a threshold (Section 8.4). Each predicted loss (vertical ticks) lies inside its measured four-fold step: the fixed sink's abstention in all four configurations, and sink $+\log N$'s exact-match retrieval, predicted at about 480 keys and lost between 256 and 1,024. The fixed sink's abstention and retrieval losses fall within one step here; at logit scale 20 with an exact match (not drawn) abstention is lost first, as in (a), at least six doublings before retrieval. The gap depends on the sink's place between background and match. With a weak match the key check outlasts tempering, losing abstention where the top background score hits its threshold. In trained 125M language models, log-N tempering did not keep abstention on top of YaRN, which already carries a log-length temperature (Section 17.4).

4. **An interpretation under a price model.** "All cheap at once" (Section 0.3.3) rests on a semi-formal price model, not on measured prices of the alternatives, and corner structure (Section 0.3.5) is a hypothesis under assumed concave prices.
5. **Retrodiction is not evidence.** The reading of the adopted changes in Section 0.3.6 is an interpretation.
6. **The test is what was stated in advance.** The account is tested by the predictions in Section 0.6.2 and the decisive tests of Section 20.4. Two of those tests, both on long-context abstention (Section 0.3.7), have run. In the first, on a production model, the outcome was undecided by its prespecified rule: one of its two corpora failed its control, and its slopes matched neither prespecified branch, though descriptively they lean toward compensation through position. In the second, in trained 125M language models, tempering every logit by $\log N / \log N_{\text{train}}$ on top of YaRN, which already carries a log-length temperature, did not keep abstention, so its lead prediction failed. Every rule that gives attention a null removed the first-token sink and most massive activations, as that test's registration predicted for language models and as earlier reports found; its registered counterpart in vision transformers did not hold, since the activation ratio rose in one seed. The others are pending.

The book develops this section in order. Section 0.3.1 is developed in Chapters 1–4; Section 0.3.2 in Chapters 5–11, one for each condition and one for the writer; Section 0.3.3 in Chapter 14; Section 0.3.4 in Chapter 15; Sections 0.3.5 and 0.3.6 in Chapter 16; Section 0.3.7 in Part V (Chapters 17–19); and Section 0.3.8 in the reading guide's account of the evidence (Section R.2), in the closing section of every chapter, and in Chapters 20 and 21.

0.4 What the question uncovered

Several of the program's main results carry the argument of Section 0.3: the six conditions, compensation, the study of explicit knobs, corners and the room. That argument is not repeated here, though some of the results it uses, such as the laws behind the room, reappear below. This section collects, in six themes, results that stand on their own for readers who do not care why transformers persist. Table 0.3 lists each with its strongest number, its evidence label and its chapter in the monograph.

0.4.1 Attention has an arrow, and default statistics of a single operator cannot see it

Many tasks need information to move one way along a sequence, and this theme asks whether attention has such a direction and whether the usual measurements can see it. Read as a Markov kernel, a head is the step rule of a random walk that moves from token to token with the head's attention weights as probabilities. Its stationary law is the long-run share of the walk's visits that each token receives, and every unmasked softmax head has a positive one. Every head with a positive stationary law splits exactly into three parts, orthogonal in the inner product weighted by that law: a shared preference for where attention gathers in the long run, a reciprocal exchange that flows equally both ways, and a circulation, a one-way current around loops of tokens that carries the head's direction (Section 3.2). At the level of pre-softmax scores, every asymmetry between token pairs splits into a key-bias potential (a score offset per key, which is what an attention sink is) plus a doubly centred cycle (with zero row and column sums), and only the cycle circulates (Section 3.3); on the complete graph of tokens this is the combinatorial Hodge decomposition used in statistical ranking (Jiang et al., 2011). This decomposition is proved, and a census applies it

to 184,320 head kernels (certified to have an accurate stationary law and time reversal) in five encoders. Learned direction is specific and consequential. The census finds that circulation peaks in the middle of the network and that the positional heads of the language encoders, which put most of their attention on the previous or the next token, circulate almost maximally. The links that touch special tokens carry a share of the circulating current 9 to 24 times their share of all links (medians over layers 1 to 10) (Section 10.3). In three vision transformers, reversing the learned circulation in every head of a middle layer changes predictions 8 to 12 times more than reversing feasible random circulations of the same size (alternatives that keep each head's stationary law and exchange fixed), as a ratio of mean changes. The prediction change from the learned reversal exceeds those circulations' mean change on at least 497 of 512 held-out images per model. An unfitted gradient readout (pairing the layer's loss gradient with the circulation update) predicts the signed change in loss under reversal with a pooled R^2 of 0.990–0.993, though its median per-image error at half dose (every head's circulation halved, once in its learned orientation and once reversed) is 8–16% (Section 10.4). The learned orientation matters, though it is not always beneficial: reversing it can lower the loss, especially in the final layer.

Default diagnostics cannot see this direction. Time reversal, which runs the head's walk backwards, flips the circulation. It leaves unchanged, though, every stationary-weighted quadratic readout of one amplified operator (an averaging filter of a single operator): a sum over tokens of a signal times its own filtered version, with each token weighted by its long-run share. Such a readout is even in the circulation: when circulation is added, scalar summaries such as collapse move only at second order, while eigenvectors move at first (Section 3.5). On per-layer lazy operators of BERT and DeiT (each head-averaged layer averaged with the identity), collapse readouts move by a median of 0.17–1.5% from their circulation-free values while eigenvectors turn by a median of 16–17 degrees (Section 3.6). Time reversal also leaves these readouts unchanged across layers that share one stationary law and form a palindromic word, a sequence that reads the same forwards and backwards. A stack of distinct layers is not a palindrome, and there collapse readouts generally move at first order too. The principle reaches beyond attention. A statistic invariant under a natural involution (an operation that, applied twice, returns what it started from), such as reversal or the exchange of two classes, can see at most the amount of the part that the involution flips, never its sign. Instruments that can see the sign must therefore be built on purpose. An associated paper on attribution gives a second instance, for the contrast between classes (Bruce, 2026f).

0.4.2 Depth is a distribution

What survives many layers of averaging depends not on the number of layers but on how often the signal takes a short route. Read a design as a path law over one fixed averaging operator: the distribution of how many averaging steps a signal takes. A mode, a pattern that one averaging step shrinks by a factor $0 < \mu < 1$ (its eigenvalue), keeps at least a fixed positive fraction of itself at every depth if and only if a fixed fraction of the paths stays within a bounded window of path lengths. Two designs with the same average depth can therefore meet opposite fates. This law is proved, and it covers oversmoothing in graph networks, the decay of label propagation and the rank collapse of pure attention repeated with one fixed operator (Sections 2.2, 2.3 and 4.1). For a stack of independent layers, how fast a mode is lost is set by the mixing budget, the expected number of layers in which the signal is averaged rather than passed on. With per-layer gains (the fraction of the mode each layer keeps) bounded away from zero, a bounded budget keeps a positive floor on what survives; the loss of the mode is polynomial when the budget grows logarithmically, stretched-exponential when it grows as a power of depth below one, and exponential when it grows linearly. To keep a floor, coherent updates (with a same-sign drift) therefore need residual branch scales of $1/L$ and incoherent

ones (centered in expectation) $1/\sqrt{L}$, or smaller, with L the total depth. In the mean-field model, which replaces each token's own normalization by its average, pre-normalized transformers start training exactly on the critical line, the logarithmic-budget boundary where loss through depth is polynomial. On this line, random, untrained networks with bidirectional attention match the predicted decay exponent at width 512, with a largest deviation of 0.041 (Sections 2.4 and 2.6).

Trained language models appear to keep their mixing budget small. Across sixteen pretrained models, on 256-token inputs, the content-mixing budget, which counts only attention that moves information to another token, not to the sink or to the token itself, stays between 1.08 and 2.81 even at 48 layers. The family slopes of budget against depth on logarithmic axes are 0.075 for Pythia, 0.376 for GPT-2 and 0.673 for OPT under norm-based shares, which set each layer's mixing share by the length of its attention update against the length of the residual stream it is added to (under energy-based shares, which compare squared lengths, Pythia and GPT-2 do not grow and OPT's slope is near one). Deeper Pythia models route their surplus attention into the sink. This is partial evidence, because the model-level intervals are wide and depth and width grow together across each family (Section 2.7).

Attention rollout (Abnar & Zuidema, 2020), a common way to trace attention through layers by multiplying residual-weighted attention matrices, turns out to have an exact meaning as a path walk. First-token collapse places all rollout mass on the first token, the causal sink. For any causal stack of stochastic attention, the largest row-wise ℓ_1 distance of rollout from this collapse is twice the worst-case probability that a walk back through the layers avoids the first token, where at each layer the walk stays put with probability equal to the residual share and otherwise steps by that layer's attention. The usual ways of handling sinks provably order the weight each content token (each token other than the sink) keeps on itself. A content token keeps the most weight on itself when the sink is treated as a skip, returning sink-directed mass to the token itself; the next most with a zero-value sink, which keeps the attention mass but writes nothing; and the least under eviction, which removes the sink and renormalizes attention over the remaining tokens. The order holds in all sixteen language models measured (Sections 2.7 and 8.3).

0.4.3 Averaging can amplify, and settling is not solving

Some models run one layer again and again on its own output, as looped and equilibrium models, iterative denoisers and test-time mixing of examples do, and their averaging rebuilds its weights from its current output at every step. Such adaptive averaging has a sharp stability test. Consider the anchored iteration $X_{t+1} = \alpha P(X_t)X_t + \beta A$ with $\alpha + \beta = 1$, which mixes the current states with weights rebuilt by the averaging rule P and pulls them back toward the fixed anchors at each step. Suppose the anchors and starting state lie in a ball of radius R, on which the rule changes by at most L_P per unit change of the state. Then a positive stability margin $m = \beta - \alpha R L_P > 0$, where the pull back to the anchors outweighs the feedback from rebuilding the weights, certifies that the ball is invariant and the map a contraction, so the iteration converges to a unique fixed point in that ball by the contraction-mapping theorem (Banach, 1922). This threshold cannot be improved over the class of such averaging rules. Tied dot-product softmax at logit scale s has $L_P \le 2sR$; the proof is formalized in Lean (Section 7.1).

Settling, though, does not mean behaving. Every equilibrium is a positive average of the anchors, which cannot amplify when its weights are held fixed (Section 0.3.1); yet because perturbing the input anchors also shifts the equilibrium weights, the equilibrium's sensitivity to input changes can approach β/m, the largest sensitivity that the stability margin m allows. Ordinary softmax kernels settle while amplifying: a Euclidean response of 18.875 falls to 1.766 when the equilibrium weights are frozen instead of adapting to the perturbation. A fixed average's response can exceed one in the

Euclidean norm, but a fixed average never enlarges the largest change at any one token; adapting weights can. A photographic denoiser amplifies a selected perturbation 19.12-fold. Test-time mixing blends each test image's frozen features with those of the other images in a context of 32 randomly drawn CIFAR-10 images before a fixed classifier reads them, much as transductive few-shot methods mix test images. Run to convergence, it cuts accuracy from 88.09% to 35.87% on test images that have the same companions under every method, whereas the same mixing with its weights frozen at the unmixed input keeps 88.02% (Sections 12.1 to 12.4). These maxima were selected from grids: they show what can happen, not how often it does.

In test-time mixing, as in any row-stochastic context mixing started at the input and followed by an affine readout, each prediction's margin between two classes splits exactly into the sample's own margin and its companions' margins for the same two classes, weighted by the mixing. This identity holds at every iteration and at the fixed point. That makes it possible to see which companions overturned a correct prediction: when strong mixing drives each sample's weight on itself down near the anchor floor, the share of itself a sample keeps once mixing has made the weights uniform (the anchor's share plus an equal part of the rest), the companions' contributions outweigh its own, and a label-free correction that subtracts the context mean recovers 87.96% (Sections 12.4 to 12.6). More generally, a computation that settles to machine precision certifies neither its response, nor the uniqueness of its endpoint, nor that the endpoint attracts, nor that a loop built on it is stable; each reading needs its own check. In a constructed two-token example, a float64 finite difference can read a response of 55.83 against a true 9,998, although the unperturbed run it starts from (the construction's clipped form) has settled to a residual of 4.97×10^{-13}. A root solver's success does not show that the root attracts nearby states. And in an outer loop, a larger loop that uses the settled result, uniformly random unit probes in n dimensions read an amplification G concentrated in one direction (rank one) as $G/\sqrt{n}$ in root mean square, so a directional probe can miss an expanding mode. These are proved and shown on constructions and small systems; trained looped and implicit models are still to be tested (Sections 13.2 to 13.6). Finally, applying the same stability margin to training ties the logit scale to weight decay. Tied dot-product softmax on a domain of radius R has loop gain at most $2sR^2$. If decoupled weight decay λ is modelled as the anchor in a training step, the learning rate scales both the feedback and the anchor pull and so cancels, giving the sufficient cap $s_c = \lambda/(2R^2)$ (Section 7.3). In the one test of this coupling so far, relaxing the anchor without any feedback reproduces all 170 flags raised by the test's divergence detector, with one false alarm among 550 cells. Because the detector fires on the anchor relaxation alone, its flags cannot test this coupling. The threshold is proved; the cap it implies is untested in language models (Section 7.4).

0.4.4 Length laws: sums, maxima, leaks and capacity

A head should be able to deliver nothing when nothing in its context is relevant, which is abstention. Every patch that lets attention abstain is the same gate on an ordinary softmax: a zero-value sink or register, an off-by-one null, a learned sink logit or a headwise output gate. With bounded scores and one null, the option that absorbs attention and delivers nothing, shared by absent and present queries (queries without or with a matching key), none keeps abstention and retrieval as the context grows. Scaling every score with the logarithm of the length is one way to keep both when the score margins between match, null and background are strictly separated (Section 0.3.7; Section 9.2).

In small byte-level language models trained without positions, a learned null, measured relative to the mean content logit (the mean score of the real keys, as against the null's), behaves as a fixed one: it does not rise with the training length. In these models, the log-odds of abstaining fall close to the fixed-null law's rate, a fall of one unit each time the length grows by a factor of e (about 2.7), across training lengths and, for models trained on 128 tokens or more, with context length within

each model (medians over heads and five seeds) (Section 17.4).

This is one instance of a broader rule. With an exchangeable background, one whose scores have the same joint distribution in any order of the items, a decision that sums over all N items at a fixed logit scale must be retuned like $\ln N$, while a decision that reads a maximum need not be retuned against backgrounds bounded below its threshold (Section 9.4). A companion bound on routing assumes independent background scores from one distribution, an independent target score, a uniformly placed target and a likelihood ratio rising with the score. Given the same scores, no router, whatever rule it uses to name an item, succeeds more often than picking the highest score. Past the evidence wall, the crowd size at which the best background key outscores the target half the time, only better keys help (Section 9.3). Both bounds are proved, with synthetic evidence; that decisions on maxima escape the schedule in trained models is partial (Section 0.3.7).

A machine that reads a store written over time often needs the latest write, not just a matching one, so it must tell later writes from earlier ones. Soft routers, readers whose weights spread over every item as a softmax's do, pay for that order as well as for the size of the crowd. To order a pair of identical writes by a ratio $r > 1$ is to give the later write at least that multiple of the earlier write's weight. With a logit range B, a router that orders its memories through a positional logit term can order at most a fraction $K/(K+1)$ of pairs of identical writes among many positions, where $K = \lfloor B/\ln r \rfloor$; a staircase positional bias attains the bound. When every position must stay retrievable, a mean-field calculation, which replaces the crowd of competing keys by its average, shows that the positional terms are paid from the router's score margin above its gain wall (their exponential mean cannot exceed the exponential of that margin), the crowd size at which the competing keys would take half of its weight. Timestamps and overwriting give order without spending that margin (Section 10.7). In the mean-field calculation of Bruce (2026e), at logit scale $\tau = 10$, with 1,024 keys in 128 dimensions and an ordering ratio of 9, the ceiling on the fraction of pairs ordered is 0.60, against 0.22 for the best linear bias.

When one context packs several documents, a head is often meant to keep each token's attention inside its own document's block. Small leaks across a block boundary grow with path length, while exact block zeros stop them. Consider a path-law filter built from one fixed averaging operator, each of whose rows leaks at most ϵ across a block boundary in one step: over a walk of random length k drawn independently of the steps, it carries at most $\epsilon\,\mathbb{E}[k]$ across that boundary. Exactly block-diagonal attention, by contrast, keeps every analytic filter block-diagonal at any depth. With logits bounded by a fixed constant, dense softmax cannot make its leak vanish as the context grows: its smallest possible leak is the routing budget's $\log N$ growth seen from the other side: holding the weight on the other blocks down as they grow takes a score range that grows like $\log N$ (Sections 6.4 and 9.1). These laws are proved, with evidence on constructed operators.

A fixed-size memory's recall limit is a capacity, not a distance: however far back the writes are, a linear memory can recall every assignment of values to N keys exactly if and only if the keys have rank N. Writes that contract the state by a common factor (a strict decay) bound the state at every length when keys and values are bounded; non-expansiveness alone does not give that bound (Sections 18.2 and 18.3).

0.4.5 Protections the transformer already has

Several proposals add explicitly what the transformer already has, such as signed attention for subtraction. Two stochastic heads, taken as arbitrary patterns of positive weights, can represent any signed kernel with constant row sums, so multi-head attention with signed output maps can subtract without signed attention weights; real heads, whose scores come from low-rank queries and keys, may not reach every pattern. In a single small block trained for 500 steps, difference

retrieval (recovering the difference of two values) under softmax attention reaches a mean R^2 of 0.46 with one head and 0.998 with four (25 matched runs at each head count). In separate six-block language and vision models, learned signed gates set the weight of a signed branch beside ordinary softmax; those weights average below one half, on the softmax side, in every run, though their final values depend on initialization: they start at 0.119 or 0.5, none of the paired starts met the registered convergence tolerance, and the language model's softmax baselines mostly missed their task threshold (Section 15.5). Positivity of the attention weights keeps outputs in the hull of the values, but it does not cap the spectral norm at one, so subtraction built from heads still needs its norm controlled. This is partial evidence: a sweep over head counts at larger scale is pending.

Invertible is not the same as isometric. An invertible block allows its input to be recovered exactly from its output, which still lets the block write features; an isometric block preserves distances between inputs, and a norm-preserving block keeps the length of each state. An invertible block built from additive reversible coupling matches a pre-normalized transformer on dictionary retrieval (mean $R^2 \geq 0.99996$ over five seeds and 8–64 associations). On CIFAR-100, it sits a mean 0.78 percentage points below that transformer (95% interval $[-1.98, 0.43]$, five paired seeds) with a fifth of the parameters. At depth 48, it needs 0.24 times the extra training memory of checkpointing, gradient buffers included. By contrast, input-conditioned orthogonal transport reaches only 0.18–0.26 on retrieval (mean R^2 at 16 associations, on a synthetic bidirectional task), and its token-rotation arm, which preserves state norms and the channel Gram matrix, loses 10 points on CIFAR-100 (Sections 11.2 to 11.5). These observations concern that transport family; the arms were not parameter-matched, and the implementations' qualification is incomplete (their reference-parity checks could not run), so this too is partial.

Production vision transformers grow sinks of their own. Six of seven pretrained ViTs grow a population of stationary-mass outliers (patch tokens that draw an outlier share of the attention chain's stationary mass) that first appear in at least half the sampled images between layers 6 and 11. These sinks do not have zero values in the six checkpoints tested for it: zeroing their values still moves the output logits, so they carry content of their own. Whether the same tokens remain sinks across layers depends on the checkpoint (Section 8.5). The census is observational: training recipe, data and scale are confounded across the checkpoints.

0.4.6 How to know what you know

A strong signal need not be useful. On fifteen neighbour graphs of frozen CIFAR-100 features (three backbones, five seeds each), each held bit-identical while the labelled set is cut from 40 to 4 examples per class, the worst tested accuracy effect of propagating predictions over the graph turns from harmful into helpful. The paired swing is 10.94 points (95% interval [10.23, 11.64]); the registered primary, pooling 4 and 8 per class, is 8.78 [8.15, 9.41]. The best spectral gauge of signal, a score computed from the graph's spectrum and the labelled nodes, rises with label density. It therefore approves propagation in exactly the high-density cases, where the worst tested effect of propagating is negative and the classifier already has little headroom, little that propagation could still add, and rejects it in the low-density ones, where that effect is positive. Utility is signal times headroom, though the label-density sweep does not separate headroom from the coverage of nodes with no labelled neighbour (Section 4.4).

Other results. A few more results across the section's themes are partial, synthetic or still unrun, and are listed here to complete the inventory. Two training loops that share state can each pass a single-loop certificate while the pair diverges. On a trained self-distillation stack, where a student and a moving-average teacher form such a pair, a single composite stability margin combining

Table 0.3: What the question uncovered: the headline results by theme, with the strongest number, the evidence label and the chapter of the monograph (Bruce, 2026a). A dagger marks an exact identity: it holds by construction for every operator of its class, and its value is what it makes measurable.

Result	Strongest number	Label	Chapter
Attention has an arrow, and default statistics of a single operator cannot see it			
Every unmasked attention head splits into preference, exchange and circulation	Applied to 184,320 certified kernels	proved†	3 and 10
Trained attention's direction is specific, and it matters	Reversal: 8–12 times the matched controls' mean change; above their mean on $\geq$ 497 of 512 images	measured	10
Standard diagnostics of a single operator cannot see which way attention points	Collapse readouts move a median 0.17–1.5% while eigenvectors turn a median 16–17°	proved	3 and 10
Depth is a distribution			
Depth is a distribution: short paths decide what survives	Retention bounds checked on 2,000 random path laws, largest violation 1.1×10^{-16}	proved	2
The expected number of mixing layers sets how fast depth forgets	Random Pre-LN transformers with bidirectional attention at initialization match the predicted exponent with a largest deviation of 0.041	proved	2
Deeper trained language models send their surplus attention to the sink	Mixing-budget slopes 0.07 (Pythia), 0.38 (GPT-2), 0.67 (OPT); wide intervals	partial	2
Attention rollout has an exact probabilistic meaning	The sink-treatment order holds for every token in 16 LMs	proved†	8
Averaging can amplify, and settling is not solving			
A sharp, machine-checked stability test for adaptive averaging	Formalized in Lean: 56 audited declarations	proved	7
A settled average can amplify changes and destroy accuracy	Matched-query accuracy 88.09% $\rightarrow$ 35.87%, with frozen weights at 88.02%	measured	12
Each prediction's class margin splits into its own and its companions'	A label-free context-mean correction recovers 87.96% (input 88.09%)	proved†	12
Converging is not the same as being right or stable	A finite difference on a settled solve reads 55.83 against a true 9,998 (unperturbed residual 4.97×10^{-13})	partial	13
One stability threshold links logit scale to weight decay	$s_c = \lambda/(2R^2)$; a feedback-free relaxation reproduces all 170 detector flags, with one false alarm in 550 cells	proved	7
Length laws: sums, maxima, leaks and capacity			
Softmax's $\log N$ law: abstention and retrieval cannot both be fixed	Fixed sink 0.874 $\rightarrow$ 0.007 at 65,536 keys, within 0.002 per seed	proved	9 and 17
Sums need a length schedule; maxima do not	The fixed softmax sink fails at the predicted onset in 4 of 4 constructed configurations	partial	8, 9 and 17
Soft routers pay for order out of their score range	Range B orders at most a fraction $K/(K+1)$ of the pairs of identical writes, $K = \lfloor B/\ln r \rfloor$; the mean-field ceiling is 0.60 at $\tau = 10$, against 0.22 for the best linear bias	proved	10
Small attention leaks can grow with depth; exact block zeros stop them	A 1% leak reaches 26.2% after 32 steps; exact zeros stay zero	proved	6
A fixed memory's recall is limited by capacity, however far back	Exact linear recall of arbitrary values needs $N \leq d_k$ per head; a trained memory holds, on 32 pairs, 0.997–0.999 across an 8,192-token gap	proved	18
Protections the transformer already has			
Signed attention is already inside your transformer	Difference retrieval R^2 0.46 with one head, 0.998 with four	partial	15
Invertible blocks keep capability; the tested orthogonal transport loses it	Retrieval $R^2 \geq 0.99996$ for additive coupling vs 0.18–0.26 for orthogonal transport	partial	11
Production vision transformers grow sinks that are not empty	Six of seven ViTs grow sinks, born at layers 6–11; not value-null	measured	8
How to know what you know			
A strong signal is not the same as a useful one	A paired swing in propagation's accuracy effect of +10.94 pp [10.23, 11.64] from 40 to 4 labels per class, on 15 graphs each held bit-identical across label densities (registered primary, pooling 4 and 8: +8.78 [8.15, 9.41])	measured	4

self-distillation with the labelled task's anchoring curvature did not certify representation collapse; a two-loop eigenvalue pencil checks their joint stability through a matrix's eigenvalues, certifying the pair and naming the coupling to cut, so far only on synthetic rigs (Section 13.8). At score collapse, when attention keys become near-equal, a fixed null competes with all of them, so only a null coupled to the temperature, or gated by the query, should repair abstention across temperatures with one shared setting, a prediction not yet run (Section 8.6). In graph-consistency regularization, holding the perturbation and the loss fixed, the neighbour graph is what carries the regularizer's improvement, on one backbone architecture so far (unreleased work). The anchored causal resolvent of a frozen, factorized kernel, which sums causal walks with a restart at the input, computes in one causal pass with a state independent of length, a design without a trained test (Section 18.1). And when attention is block-diagonal, its blocks need not share one filter: exact block zeros let each block carry its own filter with no contamination between blocks, shown on constructed operators (Section 6.4).

Each theme of this section has its chapters: Section 0.4.1 in Chapters 3 and 10; Section 0.4.2 in Chapters 2, 5 and 8; Section 0.4.3 in Chapters 7, 12 and 13; Section 0.4.4 in Chapters 6, 9, 10, 17 and 18; Section 0.4.5 in Chapters 8, 11 and 15; and Section 0.4.6 in Section 4.4, with its other results in Sections 6.4, 8.6, 13.8 and 18.1.

0.5 What holds in trained networks, and what is shown in idealized settings

Many results in the collection are proved for one operator applied again and again: a fixed matrix raised to powers, or one adaptive map iterated to its fixed point. A trained transformer is not that object. It applies a different, input-dependent operator at every layer, with residual paths, normalization and an MLP between, and it applies each layer once. So each claim this chapter quotes answers a second question beside the strength of its evidence: does the statement hold for the networks people train, or does it describe an idealized object whose behaviour trained networks may or may not share? There are five answers: a statement may hold across its stated class, hold under premises checked on the network, be shown in an idealized setting, be measured in trained networks, or be an interpretation with nothing to transfer, and Table 0.4 gives each claim's answer (along with the claims still pending).

Exact for every network of the stated class. Some statements are identities that hold at every layer and for every input, so they apply to trained networks as they stand. Every softmax head is a row-stochastic kernel whose outputs are convex combinations of its values, and every zero-valued null patch (such as a zero-value sink, register or learned sink logit) and headwise output gate factors as a gate times a softmax (Section 0.3.2) (Section 8.2). Every head with a positive stationary law splits exactly into preference, exchange and circulation (Section 0.4.1) (Section 3.2). A mixer that only averages keeps every output in the hull of its inputs and cannot raise their rank (Section 0.3.2) (Section 11.1). After row-stochastic context mixing and an affine readout, each prediction's class margin splits exactly into its own margin and its companions' margins for the same two classes, weighted by the mixing (Section 0.4.3) (Section 12.5). One result of this kind is about whole stacks: attention rollout's distance from collapse onto the first token is twice the worst-case survival probability, the probability that a walk back through the layers avoids the first token (Section 0.4.2). This holds for any ordered product of causal stochastic kernels in which the first token is absorbing, with per-token residual shares, whatever the layers (Section 8.3). It is exact for that linearization of attention flow, not for the full block.

Exact under premises checked network by network. Other statements are theorems whose conditions a given network either meets or does not, and the conditions can be measured on it. The routing budget, and the two-sided law for a null that absent and present queries share (Sections 0.3.2 and 0.3.7), hold for any softmax row whose content logits lie in $[-\tau, \tau]$, with τ read off the network (Section 9.1). The anchor exception (Section 0.3.4), which rules out compensating for a missing anchor, holds for any stack meeting these conditions: the layer maps restrict to a common discriminative invariant subspace with norms at most one; the contraction product across layers tends to zero; and the recovery map's Lipschitz gain is uniformly bounded. A version based on Dobrushin's coefficient, the factor by which one averaging step is guaranteed to shrink the spread between tokens, covers sink-like attention with shared feature maps (Section 5.5). Recall capacity (Section 0.3.7) holds for linear read-back from a memory, through the rank of its keys (Section 18.3). The parity law for stationary-weighted quadratic readouts (Section 0.4.1), which leaves them unchanged under time reversal, holds for one amplified operator with a positive stationary law and for palindromic words of layers (layer sequences that read the same forwards and backwards) that share one; a stack of distinct layers is not a palindrome, and there the readouts generally move at first order in the circulation (Section 3.5). The certificate for adaptive averaging (Sections 0.3.2 and 0.4.3) holds for any anchored iteration whose stability margin m between the anchor and the feedback is positive (Section 7.1).

Where the idealization meets practice. That last statement is the clearest case of the difference between one operator applied over and over and a different operator at each layer. The certificate, and with it the gain-control condition and the amplification results of Section 0.4.3, concern one adaptive map applied repeatedly. Weight-tied looped models, deep equilibrium models, iterative denoisers and test-time context mixing reuse the same parameters at every step. Their averaging therefore belongs to that class when it takes the anchored form and meets its premises (Section 4.2). At a fixed point, freezing the attention weights gives one fixed operator, the object the static theory describes. Linearizing the adaptive step around that fixed point adds a second term, the feedback from rebuilding the weights when the input changes, which is how a settled average can still amplify perturbations (Sections 1.4 and 12.1). Away from a fixed point even these models apply a different operator at each step, because attention depends on its input. A standard stack applies each of its distinct layers once, so it is outside the class. The per-layer feedback-gain bound $\kappa_f \le 2sR^2$, in terms of the logit scale and the representation radius (Section 0.3.2), still holds for each of its layers with tied dot-product scores, and with the projections' norms as extra factors for projected scores. But contraction and a unique attracting state are not statements about it.

Shown in an idealized setting. A third group is proved for an idealized object, and its relation to a trained network is a question rather than a theorem. The anchor's if-and-only-if condition for keeping a mode through depth (Section 0.3.2), depth as a distribution over path lengths (Section 0.4.2), and the growth of cross-block leaks with path length (Section 0.4.4) are proved for path laws over powers of one fixed operator (Sections 2.1 to 2.3 and 6.4). Linear propagation over a fixed graph satisfies this exactly; a transformer does not. What carries over is the qualitative lesson of these results. Bounds on realized products of distinct mixing kernels hold however the kernels arose (Section 2.5). The Dobrushin bound covers attention that changes from layer to layer while it mixes strictly, every layer's mixing step shrinking the spread between tokens by at least a fixed factor, with shared feature maps (Section 5.5). So averaging without short paths forgets in trained stacks too, though these bounds do not model the full block. What does not carry over is the exact criterion and its rates. The mixing-budget law (Section 0.4.2) is stated for the expected energies

of residual stacks with random feature maps and mean-field normalization, and it is matched by random, untrained transformers with bidirectional attention at initialization (Sections 2.4 and 2.6). For independent or exchangeable backgrounds on constructed keys, it is shown that routing decisions based on sums over all competitors at a fixed logit scale need a $\log N$ schedule while decisions based on maxima need none against backgrounds bounded below their threshold (Sections 0.3.7 and 0.4.4) (Section 9.4). The mean-field law that the lookup-or-abstain pilot of Section 0.3.7 tracks assumes an independent background of fixed, isotropic keys (Sections 9.2 and 17.2).

Measured in trained networks. The fourth group is measured on trained networks directly, and its reach is the reach of the models measured. Removal signatures, compensation and the study of explicit knobs (Sections 0.3.2 and 0.3.4) come from controlled runs, mostly below 25 million parameters (Sections 5.2, 15.1 and 15.4). The circulation census (Section 0.4.1) covers 184,320 head kernels of five trained encoders, with reversals tested in three vision transformers (Sections 10.3 and 10.4). Depth and sinks (Sections 0.4.2 and 0.4.5) were measured on sixteen pretrained language models and seven production vision transformers (Sections 2.7 and 8.5). For long-context abstention (Section 0.3.7), the prespecified sink test on a production language model, gpt-oss, was undecided by its rule, though descriptively its slopes lean toward compensation through position (Section 17.3). Controlled causal language models at 125 million parameters were trained at 1,024 tokens and tested to 8,192, with three seeds per rule. In these tests, log-N tempering on top of YaRN, which already carries a log-length temperature, did not keep abstention and raised long-context perplexity. Every tested rule that gives attention a null removed the first-token sink (Section 17.4). The 350-million-parameter leg is not yet measured (Section 20.4).

Interpretations. The sweet spot, the price list, the retrodiction and the successor checklist (Sections 0.3.3, 0.3.6 and 0.3.7) are readings under a model, not statements about networks, so there is nothing for them to transfer. They stand or fall with the predictions stated in advance (Section 0.6.2). The final group of Table 0.4 lists the claims whose test or theory is still pending, with nothing yet to transfer.

How to read the rest of the paper. Where Sections 0.3 and 0.4 state a law for one operator, they say so, and the practical rules in Section 0.6 carry those scopes forward. Table 0.4 is the place to check what a statement covers before applying it to a network.

Table 0.4: What holds in trained networks, claim by claim. Each claim of Sections 0.3 and 0.4 is grouped by its transfer label; beside it, what has been measured in trained networks, with the models and the scale, or the setting the statement covers when nothing has been measured. Each label's note is in the claims ledger (Appendix B). A dagger marks an exact identity: it holds by construction for every operator of its class, and its value is what it makes measurable.

Claim	In trained networks, or the setting covered
Exact for every network of the stated class	
A mixing layer is an averaging operator rebuilt from its input	Every softmax head, at every layer and input, is a row-stochastic kernel $P(X)$ whose outputs are convex combinations of its values. The path-lottery form $\mathbb{E}[P^K]$ describes powers of one operator; a trained stack is an ordered product of different $P(X)$, with MLPs between.
Attention cannot abstain unless it is given a way out	Sigmoid attention in small ViTs (3 seeds); an output gate in a small causal LM (3 seeds; read from value norms, as the stationary-mass sink test is vacuous under a causal mask); 2,700 operators captured from seven public LMs at 64 tokens.

Claim	In trained networks, or the setting covered
A head that cannot circulate cannot shift tokens around a cycle	Tied (Gram) score models reach 60% and 76% successor retrieval at depth 2 and 4; under a causal mask, with both arms masked, they copy less surely at varying offsets (mean per-seed cross-entropy ratio 3.4, accuracy 0.8 pp lower; 5 seeds); reversing learned circulation in trained retrieval models turns successor into predecessor retrieval.
Moving information around is not enough; something must write	Transport-only mixers trained on retrieval and CIFAR-100 (5 seeds; arms not parameter-matched).
Every unmasked attention head splits into preference, exchange and circulation†	Applied to 184,320 certified kernels of five trained encoders.
Attention rollout has an exact probabilistic meaning†	The ordering holds for every token in sixteen pretrained language models.
Each prediction's class margin splits into its own and its companions'†	23,040 endpoints reproduced on test-time mixing over real image contexts.
Converging is not the same as being right or stable	Constructions and small systems; trained looped and implicit models are pending.
Soft routers pay for order out of their score range	A combinatorial bound for any soft router whose positional logit term has range B; its cost to the score margin, when every position must stay retrievable, is a mean-field statement.
Signed attention is already inside your transformer	Small trained models (5 seeds); a head-count sweep at 125M–350M is pending.
Exact under premises that are checked network by network	
Routing and content need separate maps	Tied value maps at head dimension d_b in small trained models; the tying tax at depth 24 on CIFAR-100 (5 seeds).
Self-reinforcing weights need their feedback held below the anchor	Removal signature synthetic only.
One key among N needs a score margin growing like $\log N$	Accuracy onsets of small trained retrieval heads (15 cells); a training-free $\log N$ re-score on trained ViTs, on upsampled images with interpolated positions; the background's growth with position in trained causal language models at 0.6M and 125M parameters.
Every machine pays to name one of N, in some currency	All tests synthetic; no trained language model yet.
No capacity of bounded gain buys back short paths	Identity removal in trained small ViTs; the capacity sweep in trained networks is pending.
Under concave prices, the best design is a corner	Deployed hybrid fractions, recorded descriptively for two models.
No fixed rule keeps both abstention and retrieval at long context	A controlled pilot of trained lookup-or-abstain heads (fixed isotropic keys; 2 seeds, 4 configurations); the gpt-oss test was undecided by its prespecified rule (descriptively, toward compensation); in 125M language models (3 seeds) log-N tempering did not keep abstention on top of YaRN, and helped beyond the training length on interpolated positions without YaRN's temperature.
Bounded memory buys length robustness and pays in recall	A gated delta rule on multi-query associative recall; a decay gate in a trained hybrid (five seeds per variant).
Standard diagnostics of a single operator cannot see which way attention points	Native-flow readouts of lazy, head-averaged operators from two small trained encoders (BERT, DeiT).
A sharp, machine-checked stability test for adaptive averaging	A real photographic denoiser; test-time mixing over real image contexts.
One stability threshold links logit scale to weight decay	The divergence detector's anchor-relaxation confound is measured on 550 cells (all 170 flags reproduced, one false alarm); the cap is untested in language models.

Claim	In trained networks, or the setting covered
Softmax's $\log N$ law: abstention and retrieval cannot both be fixed	A controlled pilot of trained heads (2 seeds; fixed isotropic keys); small byte-level LMs without positions follow the stationary law within each model trained on 128 tokens or more; the gpt-oss test leaned toward compensation without settling it; in 125M LMs with rotary positions, abstention stays nearly level inside the training length through position.
A fixed memory's recall is limited by capacity, however far back	A gated delta rule on MQAR.
Shown in an idealized setting	
The same laws govern graphs, denoisers and attention	The realized filters on a 1,500-node connected subgraph of a citation graph, which match the closed form to 1.5×10^{-14} (the collapse-ratio identity holds to 6.5×10^{-16}); a real photographic denoiser; not a trained transformer's forward pass.
Deep averaging forgets unless some paths stay short	Removal measured in trained small ViTs (identity path removed: chance at depth ≥ 12; Post-LN cliff at depth ≥ 18 on CIFAR-10) and in weight-tied looped blocks (96 loops, 3 seeds); all below 25M parameters.
Decisions on maxima need no length schedule	Small trained heads on constructed geometries (3 and 5 seeds); in 125M language models (3 seeds), with YaRN, a head built on maxima was the only rule whose null and match weights (attention weights, not behaviour) both fell by less than half a nat, on the seed means, not in every seed; it failed natively.
Depth is a distribution: short paths decide what survives	Graph filters on real graphs.
The expected number of mixing layers sets how fast depth forgets	Matched by random, untrained transformers with bidirectional attention at initialization (width 512, largest exponent deviation 0.041).
Sums need a length schedule; maxima do not	Constructed product keys (3 seeds pooled) and small trained heads on constructed geometries (3 and 5 seeds).
Small attention leaks can grow with depth; exact block zeros stop them	Constructed operators; trained and pretrained tests are pending.
Measured in trained networks	
Trained networks buy back what a component lacks	Small trained models below 25M parameters: a rank-$(C-2)$ content map for $C = 10$ classes trained with its readout (99.84% accuracy), tied value maps, multi-head subtraction, and the tying tax at depth 24 on CIFAR-100 (5 seeds varying only the initialization).
Making a protection explicit does not beat a block that has it	Small models, mostly below 25M parameters; several planned tests remain.
A 22M-parameter accuracy margin has not yet been replicated	The minimal block's margin over ViT-Small on Tiny-ImageNet and ImageNet-100, 22M parameters, 5 seeds.
Trained attention's direction is specific, and it matters	A census of five trained encoders; reversal tested in three ViTs, one checkpoint each, on at least 497 of 512 images per model.
Deeper trained language models send their surplus attention to the sink	Sixteen pretrained language models (GPT-2, Pythia, OPT); model-level intervals are wide.
A settled average can amplify changes and destroy accuracy	Real images and frozen features; not a trained transformer's forward pass.
Invertible blocks keep capability; the tested orthogonal transport loses it	Retrieval and CIFAR-100, 5 seeds; arms not parameter-matched.
Production vision transformers grow sinks that are not empty	Seven pretrained ViTs, one checkpoint per setting; recipe, data and scale confounded.
A strong signal is not the same as a useful one	CIFAR-100 kNN graphs from three backbones, 5 seeds each; a second dataset is pending.
Interpretations	
We read the standard block as paying for every condition at once	Tuned standard components carry the protections in small trained models (Pre-LN anchor 0.998; Switch balancing; multi-head subtraction).
Hardware sets the prices; the conditions are what is priced	A reading that joins the hardware account to the conditions; there is nothing to transfer.

Claim	In trained networks, or the setting covered
Relaxing a constraint buys a capability and obliges a re-supply	A reading of the lattice; the individual re-supply obligations are proved in their chapters where stated.
Read backward, adopted changes re-supply the same protections	A reading of adopted changes; there is nothing to transfer.
Read forward, the conditions rebuild a transformer	The minimal block in small trained models: detailed balance to 10^{-12} in 50 encoders; 20/20 paired wins over ViT-Tiny on CIFAR-10/100.
What any successor must still supply	A checklist read from the stated classes; there is nothing to transfer.
Pending	
Deficits may cost less as models grow	A hypothesis about trained networks at scale; its test is pending.
What makes a good writer is still open	Open: there is no theory yet to transfer.
Predictions stated before the evidence	Predictions about trained networks, stated before their tests; the tests already run are reported with their outcomes.

Every chapter in Parts I–V closes with a section, *What holds in trained networks*, that carries these labels for its own claims.

0.6 What this changes

0.6.1 For architecture work

The account gives a way to read an architecture proposal. Ask which constraint of the block it relaxes, what that relaxation buys, and how it re-supplies what the constraint guaranteed (Section 0.3.3); a proposal that relaxes a constraint and re-supplies nothing has a failure mode waiting for it. Expect a successor to win on a resource axis, such as memory, length or speed, at matched quality, rather than on quality at matched cost. Read small-scale wins with compensation in mind: if spare capacity makes up for weakened components, an improvement measured in a small model may shrink in a large one, although that is still a hypothesis (Section 0.3.4). And test a protection by removing it and looking for its removal signature, not by adding an explicit knob for it, because a tuned block already supplies the protection (Section 0.3.4).

Some rules can be used now, each with the evidence its chapter gives. Give attention a way to deliver nothing, such as an output gate on each head or a null, an option that absorbs attention without delivering anything, like a learned sink (Section 0.3.7): in trained 125M language models both remove the first-token sink and most massive activations at no cost in perplexity at the training length (Section 17.4).

When extending a model beyond its training length, keep its positions in distribution with interpolation such as YaRN (Section 0.3.7). Without interpolation, ways to deliver nothing that act outside the attention scores, such as an output gate, degrade about as fast as plain softmax, while a fixed null in the scores, such as a sink logit, breaks down. Scaling every logit by $\log N / \log N_{\text{train}}$, the ratio of log context length to log training length, helped where positions were interpolated, in vision transformers at higher resolution (on upsampled images a constant temperature fitted per resolution did as well) and in causal language models; on top of YaRN, which already carries a log-length temperature, that scaling hurt, and without interpolation YaRN's temperature alone hurt as well (Sections 9.2, 17.4 and 17.6).

With bounded keys and values, bound a recurrent state with decay that makes every write contract by a common factor, rather than by limiting the write step alone, which makes writes non-expansive without bounding the state (Section 0.3.7). In that setting, size each head's state to the associations it must read back exactly, at most d_k of them, the key dimension, for a linear

memory (Sections 18.2 and 18.3).

Scale coherent residual branches (with a same-sign drift) by $1/L$ and incoherent ones (centered in expectation) by $1/\sqrt{L}$, with L the total depth (Section 0.4.2), as the fixed-operator energy model prescribes to keep a floor through depth (Section 2.4).

In a decoder, mask any layer built from a time reversal or a circulation flow (Section 0.4.1) and test that it mixes no future tokens into earlier outputs, because the reversal or flow of a causal operator looks ahead to future tokens (Section 10.8).

And read row-stochastic test-time context mixing with an affine readout through each prediction's own and companion margins, which split the output class margin into the sample's own contribution and the context's and show when companions overturn a correct prediction (Section 0.4.3) (Section 12.5).

0.6.2 Predictions stated in advance

The account can be wrong, and the paper released with this one on the transformer as a constrained equilibrium of protections states the predictions whose failure would show it (Section 20.1). In brief:

- The log-odds of sink mass in dense layers with learned sink logits fall with the logarithm of query position at a slope near -1, and banded layers stay flat beyond their window. The test on gpt-oss (Section 0.3.7) was undecided by its prespecified rule, though descriptively its slopes lean toward compensation through position, and in trained 125M models the fixed sinks took the compensated branch on text. Since the slope of the sink log-odds is minus the growth of each head's background, the rate at which the log of its background sum (the softmax's denominator without the sink's own term) rises with the log of the query's position, the prediction is about that growth: near one without positions, as in small byte-level models above the shortest tested training length; below one inside the training length with rotary positions, as in 125M models; and lower in more local heads. A rotary head whose background grows at the full rate inside its training length, or a local head whose background grows faster than a global one's, would count against it. This form was fitted to those models, so only other models can test it.

- Any widely adopted attention whose rows of weights need not sum to one ships with norm control. Current designs already pair the two, so the prediction concerns future designs: one adopted widely with neither output nor logit normalization would count against it.

- Sigmoid attention with a $-\log N$ bias, its offset shared by absent and present queries, loses single-key retrieval weight like $1/N$ unless its match logits grow like $\log N$. In trained 125M models with YaRN, its match weight fell from the training length to eight times it in all three seeds, by 1.14 in log weight on average against the prespecified minimum drop of 1.0 (one seed fell by only 0.48), about half the $\ln 8 \approx 2.08$ of a strict $1/N$ law; no head there met the retrieval threshold, so this concerns attention weights. The same registered prediction expected the length-calibrated sink to lose its match weight; with YaRN it lost its null mass instead (Section 17.4).

- The improvements from signed attention shrink as heads are added at fixed width.

- Masked diffusion language models carry more delivered circulation (the part of a layer's attention output that the heads' one-way current writes through their values and output maps) in early layers than masked-language-model encoders of similar size.

- Any future statistic that predicts the sign of the effect of reversing a trained head's circulation changes under time reversal (Section 0.4.1).
- Self-referential logit instabilities, in which attention's weights feed back on the tokens that set its scores, appear where the feedback outweighs the anchor, beyond the certificate (Section 0.4.3); with weight decay as the anchor and tied dot-product scores, the certificate's threshold logit scale is $\lambda/(2R^2)$, set at a fixed radius R by the weight decay alone. Because the certificate is only sufficient, this is a weak test: an onset of logit growth that moves with the learning rate at fixed weight decay would count against it.
- Sequence models keep the fraction of their layers that use full attention at 0, at 1, or near a small-denominator level (1/2, 1/3, 1/4, 1/6 or 1/8), scored by the distance to the nearest level against a null; the level set is a scoring convention, not a consequence of concave pricing (Section 0.3.5). This is the list's weakest test.
- In looped blocks without input re-injection, the fraction of a content mode that survives falls tenfold after about $2.3/(p(1-\mu_C))$ visits, with the mixing share p and the content eigenvalue μ_C measured rather than fitted.
- Mixers that only transport, mixing each group's tokens by one shared orthogonal matrix and so preserving the group's channel-by-channel inner products (its channel Gram matrix), underperform on retrieval-and-write tasks at any depth. In a provisional test, not parameter-matched, adding a writing channel restored classification but not dictionary retrieval.

Each can be checked with public models or small controlled runs.

0.6.3 Limits

Scale. The controlled evidence reaches 125 million parameters and is mostly below 25 million. Evidence at production scale is measurement on public checkpoints, and further decisive tests in trained language models remain pending (Section 20.4).

The price model. Calling the protections "cheap" (Section 0.3.3) is a judgment under a semi-formal price model, not a computation. We did not measure the alternatives' prices on a common budget of hardware and training, and the prices themselves co-evolved with the transformer as hardware, kernels and recipes adapted to it, so the sweet spot is a reading under that model, not a measured optimum.

Retrodiction. Reading past design changes through the conditions (Section 0.3.6) is an interpretation; only predictions made in advance (Section 0.6.2) can test it.

Decoders and the MLP. The vocabulary of stationary laws, reversal and circulation (Section 0.4.1) needs a positive stationary law, which causal softmax attention lacks. Causal softmax attention takes its direction from the mask. For decoders we use only statements proved for them, such as the absorbing first token and rollout as survival (Section 0.4.2) and the two-sided law (Section 0.3.7). Separately, there is no theory here of what the MLP should compute, beyond that it writes (Section 0.3.2) and that it is the main resource that compensation draws on (Section 0.3.4).

A conditional inventory. The conditions hold within their declared classes (Section 0.3.2) and say nothing outside them. Consider how content survives depth. For normalized nonnegative filters of a fixed mode $0 < \mu < 1$, content survives depth in only one way, by keeping short-path mass. The two other ways, signed coefficients and invertible transport (Section 0.4.5), work by leaving that class. Without the fixed-mode premise, even this list of three is incomplete: an operator that moves toward the identity as depth grows keeps content with none of the three, because the operator itself changes with depth (Section 14.2).

Idealized settings. Many results are proved for one operator applied repeatedly; Section 0.5 says which, and what has been measured in trained networks.

0.6.4 Components and roots

Most protections in use were found to be needed when a component broke: pre-normalization when deep post-norm models would not train (Wang et al., 2019), normalized queries and keys when attention logits grew at scale (Dehghani et al., 2023), sink tokens when evicting the first tokens broke streaming generation (Xiao et al., 2024), registers when vision transformers showed high-norm artifact tokens (Darcet et al., 2024), log-N scaling and temperature when models failed beyond their training length (Su, 2021; Chiang & Cholak, 2022; Peng et al., 2024), and hybrids when pure recurrences struggled to recall (Arora et al., 2024a,b). Component work built the transformer, but it does not by itself say why most modifications barely help, which parts are essential, or where to look next. Asking what each component *protects* does so, and in this collection that question kept proving fruitful. Asking why sinks appear led to one common form, a gate times a softmax, for every zero-valued null patch and headwise output gate, and to the two-sided law. Asking why the residual cannot be removed led to path laws and to the anchor exception. The missing anchor is a deficit that no capacity of bounded gain (bounded sensitivity) can make up for under strict mixing, the only one so proved; in the small-model ablations nothing else in the block made up for it, and a test that adds capacity is pending. Asking whether the value map matters led to compensation. Asking what bounds attention's loop gain led to a certificate, and to a test showing that divergence flags can mistake the anchor's relaxation for feedback. Asking whether an asymmetric attention matrix is directional led to circulation and the parity law. These two kinds of work, fixing components and asking what they protect, are complementary, and the second is slow. Its negative results count as evidence: the results for explicit protection knobs at small scale (Section 0.3.4) support the reading that a tuned standard block already supplies its protections, and compensation was found because necessity predictions kept failing in trained networks. The transformer was built by fixing what broke; asking what each fix protected, and keeping what failed along the way, is how we began to learn what holds it in place.

This project's ambition has outlasted my endurance. An account this broad is surely wrong in places, partly because it is this broad, so I would ask readers to poke and prod at it; the predictions stated in advance (Section 0.6.2) and the tests still pending (Section 20.4) are the quickest ways to find where. Part of why the transformer abides is fast matrix multiplication, engineering and history; part, I have argued, is that it pays cheaply for what attention needs. That explains its past, not its future. Transformers will keep changing where room remains, and an architecture that needs none of this, or buys it more cheaply, need not look anything like one.

Chapter R

Reading guide

R.1 Where each part of the argument is developed

This section maps the argument of Chapter 0 onto the chapters that develop it. Chapter 0 states the argument in brief; the chapters give each part its full statement, its evidence and its limits.

Table R.1 sets each section of Chapter 0 beside the chapters that develop it. It is the map that Chapter 0's pointers give section by section, in one table.

Table R.1: From Chapter 0 to the chapters. Each row is a section of Chapter 0; the right column names the chapters and sections that develop it.

In Chapter 0	Developed in
A sweet spot of conditions	
Section 0.3.1, the object	Chapters 1–4
Section 0.3.2, six conditions	Chapters 5–10, one per condition, and Chapter 11, the writer
Section 0.3.3, the sweet spot	Chapter 14
Section 0.3.4, a flat neighbourhood	Chapter 15
Sections 0.3.5 and 0.3.6, corners and rediscovery	Chapter 16
Section 0.3.7, the room	Chapters 17–19 (Part V)
Section 0.3.8, how strong the case is	Section R.2; the closing section of every chapter; Chapters 20 and 21
What the question uncovered	
Section 0.4.1, the arrow	Chapters 3 and 10
Section 0.4.2, depth	Chapters 2, 5 and 8
Section 0.4.3, feedback	Chapters 7, 12 and 13
Section 0.4.4, length laws	Chapters 6, 9, 10, 17 and 18
Section 0.4.5, protections already present	Chapters 8, 11 and 15
Section 0.4.6, method	Section 4.4; its other results in Sections 6.4, 8.6, 13.8 and 18.1
What this changes	
Sections 0.6.1 and 0.6.2, rules and predictions	Chapter 20, and the rules in Section 21.1 and Section 10.8
Sections 0.6.3 and 0.6.4, limits and close	Chapter 21

Most readers will want part of the book rather than all of it. The paths below cut across Table R.1 by topic. Each is a few chapters, read in order, with the papers whose results they

develop.

Long context. Chapter 8 shows why weights that sum to one leave a head no way to attend less, and that every zero-valued null patch or headwise output gate factors as a gate times a softmax. Chapter 9 adds the routing budget, a margin that grows with length, and Chapter 17 puts abstention and retrieval together at long context, from a controlled pilot to a production model and trained models up to 125M parameters, with the 350M leg still to run. Chapter 19 prices the same conditions in other machines' currencies, and Section 6.4 shows how cross-block leaks can grow with path length, while exact block separation keeps them zero (Bruce, 2026c,d,e).

Recurrent and hybrid models. Chapter 14 shows how relaxing one of the transformer's constraints obliges a protection. Chapter 18 follows recurrent and streaming mixers through that lattice: a common contraction factor keeps state bounded when keys and values are bounded, a linear memory's per-head capacity limits exact recall of arbitrary associations however far back an item lies, and hybrids keep a reserve of attention layers. Chapter 19 compares the currencies in which machines pay, and Section 20.4 registers the tests still open, among them a sweep of the fraction of attention layers in a hybrid (Bruce, 2026d,e).

Depth, residuals and normalization. Chapter 1 reads one averaging operator as a walk, and Chapter 2 shows that depth is a distribution: for one fixed operator, what survives many layers depends on how often the signal takes a short route, the mixing budget sets how fast the rest is lost, and, at initialization, normalization placement is a choice of that budget. Chapter 5 gives the anchor, its removal signatures and the one deficit that capacity elsewhere has not bought back, and under strict mixing cannot at bounded gain. Chapter 15 shows that a small trained network can evade a necessity theorem when another resource has room and its retained state still separates the task's outputs, and Sections 16.6 and 16.7 assemble the smallest block that meets the conditions, with the evidence recorded for it and the tests still to run (Bruce, 2026k,i,d).

Stability, fixed points and looped models. Chapter 7 holds feedback below the anchor of Chapter 5: it proves the certificate, computes softmax's feedback constant and reads weight decay as an anchor. Chapter 12 shows that a converged average can still amplify changes in its inputs and overturn the predictions read from it. Chapter 13 shows that a computation which settles to machine precision certifies neither its response, nor its uniqueness, nor that it attracts, nor that a loop built on it is stable, and gathers a battery of checks for fixed-point computations. These results are proved and shown on constructions and small systems; trained looped and implicit models are still to be tested (Bruce, 2026h).

Direction. Chapter 3 splits every head with a positive stationary law into preference, exchange and circulation, locates direction exactly, and shows why default statistics cannot see it. Chapter 10 takes direction up as a condition, since a head without circulation cannot shift tokens around a cycle, and routes reliably along an open sequence only by piling its stationary mass toward one end: it measures circulation in pretrained models, shows that the learned orientation is consequential, and gives the price of order (Bruce, 2026g,i,e).

Evidence and method. The account of the evidence in Section R.2 comes first. Section 15.4 reports a preregistered study in which a protection made into an explicit knob did not beat the block's own mechanism. Section 4.4 asks when propagation helps frozen features, since a strong signal need not be useful. Chapter 20 gives the predictions stated in advance and the register of open tests, and Appendix B the evidence ledger.

Figure R.1 shows how the collection's papers, whose results make up the chapters, build on one another.

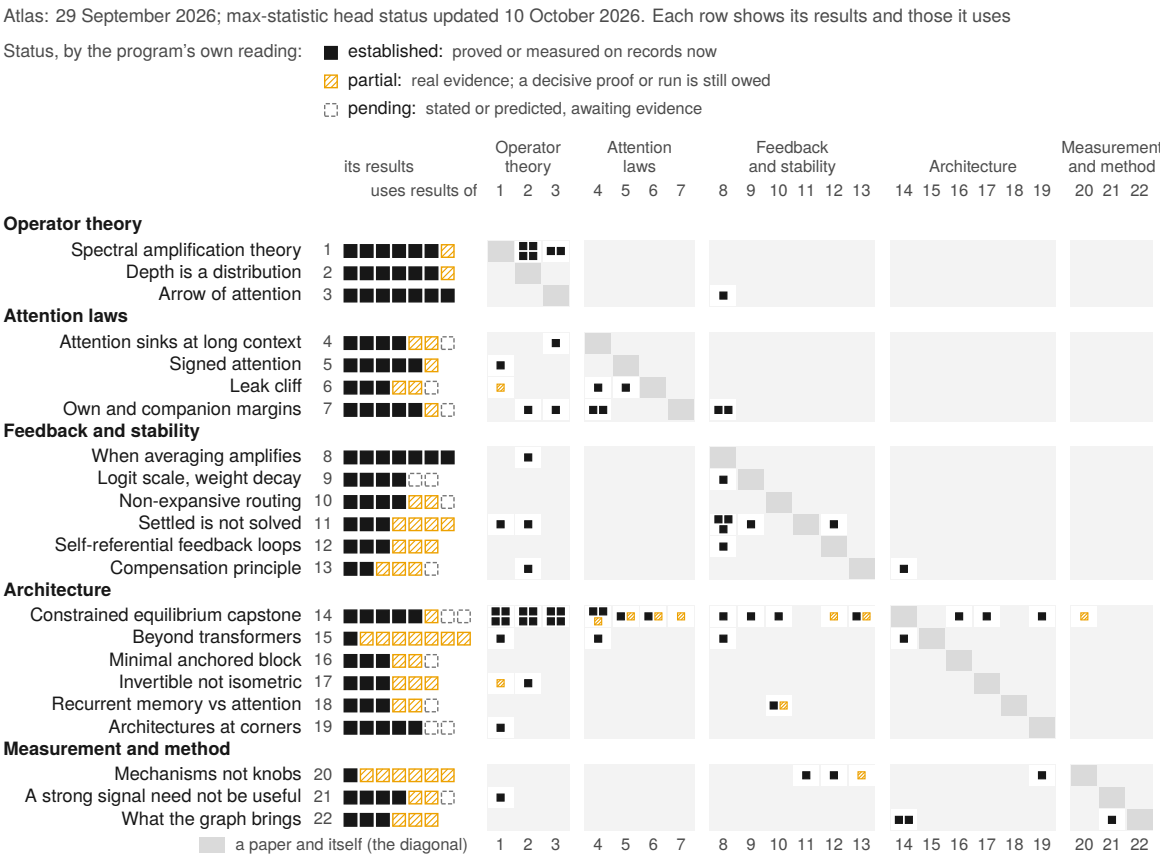


Figure R.1: **The collection's 22 papers build on one another, mostly on results that are, by the program's own reading, proved or measured: 86% of the 77 uses, against 59% of all results.** From the contribution atlas, which maps the papers' results and dependencies; most of the 22 papers are not yet released. Each row is a paper: its own results, one square each (six to eight per paper, as the atlas counts them, not a measure of size or weight), then, in the column of each other paper, the results of that paper it uses, one square each, at their own status. A paper uses a result when one of its own results relies on it (uses, generalizes or specializes it), as each paper's plan states or clearly implies, matched to results by the atlas; a use is not a citation. Rows and columns follow the atlas's families and order; the numbers are labels, not a ranking or a reading order. A status says how a result is supported, not whether it holds in trained networks (Section 0.5), nor does it pass along a use: three established results rely on a partial one. Every other family uses results of operator theory; the architecture capstone (Bruce, 2026d), built to gather the collection, uses 16 of the other 21 papers, and 7 of the 11 uses of partial results are its; no use is of a pending result. The statuses are the atlas's own, as dated in its header, and predate later results: the gpt-oss sink test, drawn as pending, has since returned, undecided by its prespecified rule (Section 17.3), and the 125M harness has reported on tempering (Section 17.4). The atlas's *established* groups the evidence labels proved and measured of Section R.2, and it has no status for a failed prediction. The results of paper 22 are not yet printed in this book (Section 4.5). Not drawn: 12 tests and 7 corrections of another paper's results, 5 uses planned only by pending results, dependencies within a paper (74), 36 links between papers that share data or runs, and the atlas's two papers outside the collection.

R.2 How to read the evidence

Every claim of Chapter 0 carries two labels. Its evidence label says what kind of support it has: *proved*, *measured*, *partial*, *pending*, *interpretation*, or *conventional*, the last for an answer from the literature that the collection does not test. Its transfer label says what carries over to trained networks: exact for every network of the stated class; exact under premises checked network by network; shown in an idealized setting; measured in trained networks; an interpretation, with nothing to transfer; an explanation from the literature; or pending, with nothing yet to transfer. Section 0.5 explains the transfer labels, and each chapter's closing section applies them to its own main claims; Appendix B lists each claim of Chapter 0 with both labels and the sections that develop it. A test is *registered*, or *prespecified*, when its predictions, analysis and decision rule were fixed, with checksums, before its run; none was publicly deposited. Where the registration itself stayed a draft, as for the production-model test of Section 17.3, the test is called prespecified, not registered. A pass of an implementation gate is not, by itself, support: a cell can pass every correctness check while the effect it was built to measure is absent. A pass supports only what the test measured, and the evidence label follows the data, not the verdict.

R.3 How the chapters depend on each other

Part I supplies the tools: the operator and its lotteries, path laws, the three-way split and parity, and the same laws in other domains. Part II states the conditions with them. Part III extends Chapter 7. Part IV reads the transformer block against Parts I and II, and Part V names the room that remains.

The prerequisites of the chapters in Parts II–V: Chapter 5 needs Chapter 2; Chapter 6 needs Sections 1.2 and 4.3; Chapter 7 needs Section 1.4; Chapter 8 needs Section 1.1; Chapter 9 needs Chapter 8; Chapter 10 needs Chapter 3; Chapter 11 needs Section 1.1; Chapters 12 and 13 need Chapter 7, and Chapter 13 also needs Chapter 12; Chapter 14 needs Part II; Chapter 15 needs Section 6.6 and Chapter 14; Chapter 16 needs Chapter 14; Chapter 17 needs Chapters 8 and 9; Chapter 18 needs Chapter 14 and Section 9.5; Chapter 19 needs Chapter 14 and Sections 9.3 and 9.5; and Chapter 20 needs Parts II and IV and Chapters 17 to 19.

R.4 Notation

The notation table is Appendix C: every symbol that recurs, with its meaning and the chapters that use it. This section keeps only the choices a reader meets first.

The sources write some objects differently, and the book fixes one name for each. The reversible part of a walk is S, its time reversal $P^\dagger$ and its circulation J, also called its flow (hence the flow ratio Φ), so $P = S + J$. Two quantities that the sources both call κ, and a third that they write κ_{f}, are told apart: κ_{col} is the collapse ratio (Chapter 2), κ_{circ} the circulation index (Chapter 3) and κ_{f} the feedback constant (Chapter 7). The two orders of amplification and a second map output Y_{pre} and Y_{post} (Chapter 6). The stationary flow $D_\pi P$ splits into the reciprocal traffic A and half the current C, $D_\pi P = A + \frac{1}{2}C$ (Chapter 3), which one source writes as a symmetric part S (the traffic) and an antisymmetric part A (half the current). Some letters keep more than one meaning, each defined where it is used, and Appendix C lists every meaning that recurs, with the chapters that use it. The logit scale is s in some sections and τ in others, as their sources write it; τ is also a bound on the logits, a temperature, an absorption time and Dobrushin's coefficient. μ is an eigenvalue, and a background mean in Chapter 9; R is a radius, a resolvent, a ratio, a dynamic range and the width

of a logit window; β is both a content gap and an anchor's share; and γ is a spectral gap, a retention level, a feedback strength and, in Chapters 9, 17 and 20, the growth of a head's background. An update's Jacobian is DF; some reprinted figures write it J, and their captions say so.

Part I

The object

Chapter 1

An operator read as a walk

Chapter 0 describes the object in words (Section 0.3.1): a token-mixing layer is an averaging operator, rebuilt from its own input and applied again and again through depth. This chapter defines it. The premise behind it, that attention's bias is not weak but averaging, is stated in Section 0.2.5 and is not repeated here.

1.1 The object: an averaging kernel rebuilt from its input

A graph-convolution layer propagates features with a normalized adjacency, a kernel method smooths with a row-normalized affinity, and an attention head mixes tokens with a row-softmax of similarity scores. Strip away the learned parts and the same object remains three times:

$$P = D^{-1}A \ \text{(graphs)}, \qquad P = D^{-1}W \ \text{(kernels)}, \qquad P = \operatorname{softmax}\left([\, q_i^\top k_j \,]_{ij}/\sqrt{d_k}\right) \ \text{(attention)}, \tag{1.1}$$

with D the diagonal matrix of row sums, queries q_i and keys k_j in $\mathbb{R}^{d_k}$: a matrix with nonnegative entries whose rows sum to one. Row i is a probability distribution (where token i looks), so P is one step of a random walk on the N tokens, and PX replaces each token's feature vector by the walk's one-step average. (The symmetric normalization $D^{-1/2}AD^{-1/2}$ is a diagonal change of coordinates away from $D^{-1}A$ and has the same spectrum.) Much of what follows concerns what happens when this step is amplified, by depth, by a restart or by a residual stream, and then read out.

For token representations $X \in \mathbb{R}^{N\times d}$, a softmax head computes

$$P(X) = \operatorname{softmax}_{\text{row}}\left(XW_QW_K^\top X^\top/\sqrt{d_k}\right)$$

and delivers $P(X)XW_VW_O$. The kernel is row-stochastic, $P\mathbf{1} = \mathbf{1}$: one step of the walk above. When it is irreducible, as full softmax attention is, it has a unique positive stationary law π. In the geometry $\langle u, v\rangle_\pi = \sum_i \pi_i u_i v_i$ it has norm one, attained on the constant (Perron) mode, and so it cannot amplify. A causal mask makes the first token absorbing and leaves no positive stationary law, so decoders are read in a different norm, and the reversible vocabulary of Section 1.2 and Chapter 3 is not used for them (Section 2.5).

Two extensions of the object are taken up later: kernels whose rows may sum to less than one, which let a head abstain (Chapter 8), and differences of kernels with signed weights, whose outputs can leave the hull of the values (Section 15.5).

Because the kernel is rebuilt from its input, the layer has feedback (Section 1.4, Chapter 7); recurrent mixers keep a matrix state instead, whose bounds are the subject of Chapter 18.

A pre-normalized residual block adds to the kernel the identity path (the anchor), normalization (radius control), per-head value and output maps (the content channel, and subtraction across

heads), and an MLP (the feature writer). These are the parts that keep the averaging operator from its failure modes. That is why the usual answer that attention succeeds by assuming little about which inputs interact is corrected in Section 0.2.5: attention's bias is averaging, and the rest of the block can be read as protection against the failure modes of averaging, an interpretation that Chapter 14 develops (Section 14.5).

1.2 The walk's own geometry

Every walk has a *stationary distribution*, a probability vector π with $\pi^\top P = \pi^\top$. If the walk is *irreducible* (every token can reach every other, which is automatic for full softmax attention), π is unique and positive (Section A.1, T5). The stationary distribution is the ruler the operator hands us: write

$$\langle u,\, v\rangle_\pi = \sum_i \pi_i\, u_i v_i, \qquad \|u\|_\pi = \langle u,\, u\rangle_\pi^{1/2},$$

with a conjugate on the first slot for complex vectors, and $\|M\|_\pi$ for the operator norm this induces. The vector $\mathbf{1}$ is fixed by P; the π-orthogonal complement of $\mathbf{1}$, the part of a signal that averaging can destroy, we call its *content*. For feature matrices, $\|X\|_\pi^2 = \sum_i \pi_i \|x_i\|^2$ row by row.

The adjoint of a matrix in this geometry is $M^\dagger = D_\pi^{-1} M^\top D_\pi$ with $D_\pi = \mathrm{diag}(\pi)$, so that $\langle u,\, Mv\rangle_\pi = \langle M^\dagger u,\, v\rangle_\pi$. Applied to the walk, $P^\dagger$ has entries $\pi_j P_{ji}/\pi_i \geq 0$ and rows summing to one (by stationarity), and it has the same stationary distribution: $P^\dagger$ is the *time reversal* of P, the walk watched backwards. The walk is *reversible* (detailed balance, $\pi_i P_{ij} = \pi_j P_{ji}$) exactly when $P^\dagger = P$. Row-normalizing a symmetric affinity always gives a reversible walk, with π proportional to the row sums: graph convolution on an undirected graph, a symmetric kernel smoother, and an attention head with symmetric scores and no mask (for instance, queries and keys from one map) are all reversible. A head with positive semidefinite scores, such as the Gram scores of tied queries and keys, is more than reversible:

Proposition 1.1 (Tied positive-semidefinite scores)**.** *If the score matrix s of a head is positive semidefinite, for instance a tied Gram form $s_{ij} = \langle Wx_i, Wx_j\rangle$, the kernel $K = \exp(s)$, taken entrywise, is positive semidefinite, and $P = D^{-1}K$ with $D = \mathrm{diag}(K\mathbf{1})$ is a reversible stochastic matrix whose spectrum lies in $[0, 1]$.*

The proof is in Section A.2. Trained attention, by contrast, is directed, because its scores are asymmetric or masked.

Every walk with a positive stationary distribution splits uniquely into a π-self-adjoint part, itself a reversible walk, and a π-antisymmetric remainder,

$$P = S + J, \qquad S = \tfrac{1}{2}(P + P^\dagger), \qquad J = \tfrac{1}{2}(P - P^\dagger), \tag{1.2}$$

where S is itself a reversible walk with the same π (the *grouping*: who averages with whom) and J is π-antisymmetric, $J^\dagger = -J$, with $J\mathbf{1} = 0$ and $\pi^\top J = 0$ (the *routing*: circulation that moves mass around loops without changing where the walk lives) (Fill, 1991); Section 3.2 develops the split. It is weighted by π: the plain symmetrization $(P + P^\top)/2$ is a different object. One family of operators organizes Sections 3.5 and 3.6: for $t \in [-1, 1]$,

$$P_t \;:=\; S + tJ \;=\; \tfrac{1+t}{2}\, P + \tfrac{1-t}{2}\, P^\dagger, \tag{1.3}$$

a convex combination of two walks with stationary distribution π, hence itself such a walk; $P_1 = P$, $P_0 = S$ and $P_{-1} = P^\dagger$. Changing the sign of t reverses time.

Lemma 1.2 (Conservation and contraction)**.** *Let P be row-stochastic with stationary distribution $\pi > 0$.*

(i) $\langle x, Jx\rangle_\pi = 0$ *for every real vector x: routing contributes nothing to any real quadratic form.*

(ii) $\|P\|_\pi = \|P^\dagger\|_\pi = \|S\|_\pi = 1$*; hence $\|P^n\|_\pi = 1$ for every n, and the* flow ratio $\Phi := \|J\|_\pi$ *lies in $[0, 1]$, the value 1 being attained (by the directed 4-cycle).*

(iii) *S has real eigenvalues $1 = \mu_1 \geq \mu_2 \geq \cdots \geq \mu_N$ in $[-1, 1]$ with a π-orthonormal eigenbasis $\{v_k\}$, $v_1 = \mathbf{1}$. If P is* lazy *($P_{ii} \geq \frac{1}{2}$ for all i), every $\mu_k \in [0, 1]$.*

The proof is in Section A.2. Part (ii) rules out transient growth in the walk's own metric: non-normal matrices can amplify perturbations transiently even when all their eigenvalues are small (Trefethen & Embree, 2005), but a row-stochastic operator measured in its own π-geometry cannot. The price of directedness is therefore not an instability; it is the error with which statements proved for reversible operators transfer to directed ones, which Chapter 3 computes. The flow ratio has a geometric reading: Φ is the half-height of the operator's numerical range in the π-geometry, whose real extent is fixed by S alone (Proposition A.2). The hypothesis $\pi > 0$ matters: strictly causal attention violates it, and Section 2.5 replaces the π-geometry there. Where eigenvalues in $[0, 1]$ are needed we use the lazy operator $\frac{1}{2}(I + P)$, which is the residual block with update share $\frac{1}{2}$; it maps each eigenvalue μ of S to $\frac{1}{2}(1 + \mu)$, and so halves both the flow ratio and the gaps. No qualitative statement changes, but every quantity must be reported with the form it was measured in.

1.3 Amplifiers are lotteries over path lengths

Depth, restarts and residual streams amplify the walk by taking nonnegative combinations of its powers.

Definition 1.3 (Amplifier, lottery, anchor)**.** An *amplifier* is a power series $g(z) = \sum_{k\geq 0} c_k z^k$ with $c_k \geq 0$ and $0 < g(1) < \infty$. Its *lottery* is the probability distribution $p_k = c_k/g(1)$ of a random path length K, and its normalized form is

$$f(z) = \frac{g(z)}{g(1)} = \mathbb{E}[z^K], \qquad f(P) = \mathbb{E}[P^K] = \sum_k p_k P^k.$$

The *anchor* is $a = \Pr[K = 0] = f(0)$, the share of the output that is the unpropagated input.

The series converges for every row-stochastic P, since every power is row-stochastic. Four lotteries account for most of practice (Table 1.1). Depth L is the deterministic lottery $K = L$. A restart with continuation probability α (personalized PageRank, label spreading, APPNP's teleport) is the geometric lottery $\Pr[K = k] = (1 - \alpha)\alpha^k$: at each step, continue with probability α or return home. A residual stack is a binomial lottery, and a residual stream whose total update stays fixed as depth grows is, in the limit, a Poisson lottery. The nonnegativity of the coefficients is not a convenience; it characterizes the filters that are themselves averaging operators.

Proposition 1.4 (Averaging filters are exactly the nonnegative ones)**.** *Let $f(z) = \sum_{k=0}^{d} c_k z^k$ with $\sum_k c_k = 1$. Then $f(P)$ is row-stochastic for every row-stochastic P if and only if $c_k \geq 0$ for all k.*

Proof. If all $c_k \geq 0$, $f(P)$ is a convex combination of row-stochastic matrices. Conversely, let Z be the cyclic shift on $N > d$ tokens. The matrices $Z^0, \ldots, Z^d$ have disjoint supports, so the entry of $f(Z)$ in row 1, column $1 + k$ is exactly c_k; if $f(Z)$ is to be nonnegative, every $c_k \geq 0$. □

Table 1.1: Four amplifier species as lotteries. Collapse ratios are for a content gap β (Section 2.3); the last column decides their second-order behaviour (Proposition 2.10): positive for under-dispersed lotteries, negative for over-dispersed ones. The restart has mean $\bar{k} = \alpha/(1-\alpha)$.

Filter	$f(\mu)$	**Lottery** K	**Anchor** a	$\kappa_{\mathrm{col}} = 1/\mathbb{E}[(1-\beta)^K]$	$\mathbb{E}K - \operatorname{Var} K$
Depth L	μ^L	$K = L$	0	$(1-\beta)^{-L}$	L
Restart α	$\dfrac{1-\alpha}{1-\alpha\mu}$	geometric	$1-\alpha$	$1 + \dfrac{\alpha\beta}{1-\alpha}$	$-\bar{k}^2$
Residual stack, share b, depth L	$(1-b+b\mu)^L$	binomial(L, b)	$(1-b)^L$	$(1-b\beta)^{-L}$	Lb^2
Residual stream, total update σ	$e^{-\sigma(1-\mu)}$	Poisson(σ)	$e^{-\sigma}$	$e^{\sigma\beta}$	0

A filter with a negative coefficient is therefore not an average over walks for every operator, and it can do things no averaging filter can: Section 2.3 shows that such filters escape the characterization of bounded collapse, and this proposition says what they give up.

Residual streams fit the same mould, one layer at a time.

Proposition 1.5 (Residual streams are lotteries over layers)**.** *Let $P_1, \dots, P_L$ be row-stochastic and $b_1, \dots, b_L \in [0,1]$. Then*

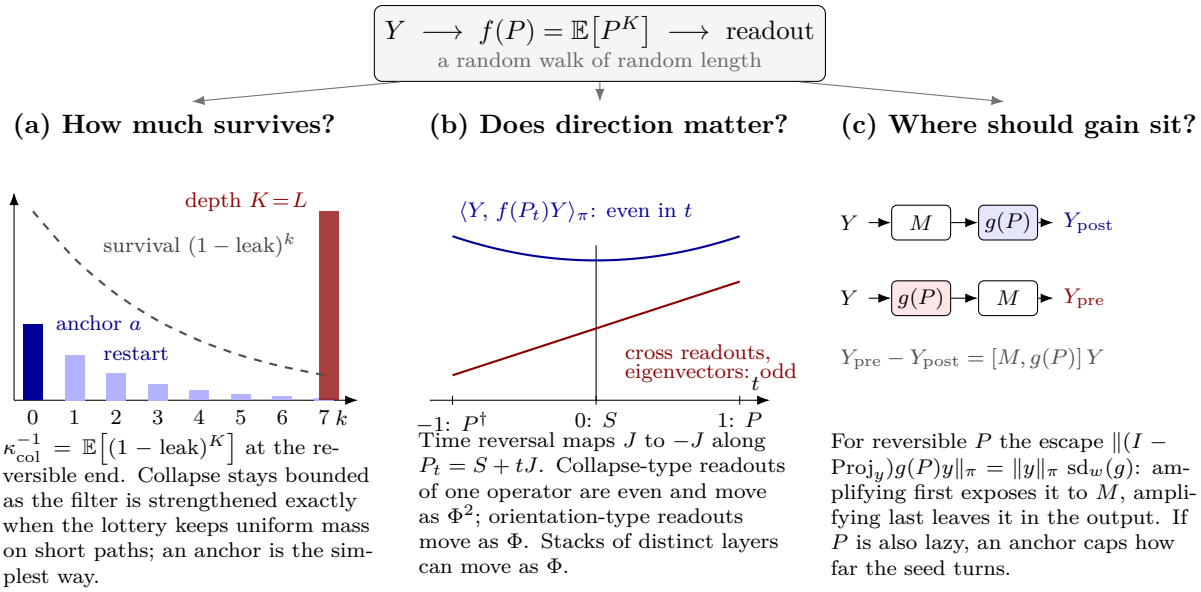
$$\prod_{\ell=L}^{1} \left((1-b_\ell)I + b_\ell P_\ell\right) \;=\; \mathbb{E}\Big[\overleftarrow{\prod_{\ell\in\mathcal{S}}} P_\ell\Big],$$

where the random subset $\mathcal{S} \subseteq \{1, \dots, L\}$ contains each layer independently with probability b_ℓ and the product runs in decreasing order of ℓ. The anchor is $\Pr[\mathcal{S} = \emptyset] = \prod_\ell (1-b_\ell)$. For a repeated operator ($P_\ell = P$, $b_\ell = b$) the path length $|\mathcal{S}|$ is binomial(L, b); with $b = \sigma/(L+\sigma)$, which holds the total update Lb below σ, the stack converges to the Poisson lottery $e^{\sigma(P-I)} = \mathbb{E}[P^K]$, $K \sim$ Poisson(σ), with $\|\cdot\|_\infty$-error at most $4\sigma^2/L$.

Proof. Expand the product: each factor contributes I (weight $1-b_\ell$) or P_ℓ (weight b_ℓ), and the choices are independent. The limit follows from Le Cam's Poisson approximation of the binomial path length, given in Section A.2. □

The proposition is the token-side half of the path expansion of a residual network (Elhage et al., 2021; Veit et al., 2016). In a block $X \mapsto X + AXW$ the attention A acts on the rows (tokens) and the learned maps act on the columns (features), so the expansion is a sum over layer subsets with attention products composing on the left and feature products on the right. With a scalar update share, $X \mapsto (1-b_\ell)X + b_\ell A_\ell X$, the token side is exactly the lottery of Proposition 1.5, and its $\mathcal{S} = \emptyset$ term, the input itself, is the anchor. (An unnormalized block $I + bA$ equals $(1+b)\big((1-b')I + b'A\big)$ with $b' = b/(1+b)$, and no statement below depends on the scalar $1+b$.) With learned value and output maps the path weights are matrices rather than probabilities, and the lottery reading describes the token side only. Collapse itself happens on this token side. Propagation contracts the π-distance of a feature matrix to the consensus set $\{\mathbf{1}c^\top\}$ by the factor $1-\gamma$ per step for a lazy reversible operator with spectral gap γ; a feature map can rescale that distance by at most its norm; and a pointwise 1-Lipschitz nonlinearity cannot expand it (Proposition A.1 in Section A.2; the argument is that of Oono & Suzuki (2020) and Cai & Wang (2020)). Learned feature maps can slow collapse only by growing their norm. The theory therefore studies the token side alone.

Figure 1.1 gathers the object and the three questions that the next chapters ask of it.



$\kappa_{\text{col}}^{-1} = \mathbb{E}[(1 - \text{leak})^K]$ at the reversible end. Collapse stays bounded as the filter is strengthened exactly when the lottery keeps uniform mass on short paths; an anchor is the simplest way.

Time reversal maps J to $-J$ along $P_t = S + tJ$. Collapse-type readouts of one operator are even and move as Φ^2; orientation-type readouts move as Φ. Stacks of distinct layers can move as Φ.

For reversible P the escape $\|(I - \text{Proj}_y)g(P)y\|_\pi = \|y\|_\pi \, \text{sd}_w(g)$: amplifying first exposes it to M, amplifying last leaves it in the output. If P is also lazy, an anchor caps how far the seed turns.

Figure 1.1: **One object, three questions.** A row-stochastic operator amplified by nonnegative path weights is a random walk of random length. (a) Its collapse is a race between the path-length lottery and a leak of the operator (Section 2.3); a restart keeps an anchor at $K = 0$, while depth puts all its mass at $K = L$. (b) Reversing time flips the flow, so quadratic readouts of one operator are even in it (Section 3.5). (c) The two orders of amplification and a second map differ by a commutator; for a reversible operator its worst case is the gain's spread over the seed's spectrum (Sections 4.3 and 6.2).

1.4 Adaptive mixing: the iteration Parts II and III study

Chapter 7 (in Part II) and Part III study the layer as an iteration. Let $A \in \mathbb{R}^{N \times d}$ contain fixed input representations, called *anchors*, and X_t the evolving states. A *builder* $\mathcal{P}(X)$ returns nonnegative weights whose rows sum to one, and the iteration is

$$X_{t+1} = F_A(X_t) = \alpha \mathcal{P}(X_t) X_t + \beta A, \qquad \beta = 1 - \alpha, \quad 0 \le \alpha < 1. \tag{1.4}$$

In RED denoising with a non-local-means denoiser, A is the observed image, X_t its current estimate, and $\mathcal{P}(X_t)$ the row-normalized non-local-means weights; the published fixed-point rule with identity observation operator is exactly (1.4), with $\alpha = \lambda\sigma^2/(1 + \lambda\sigma^2)$ for regularization weight λ and observation-noise variance σ^2 (Romano et al., 2017). Fixed-graph propagation such as APPNP is, after degree rescaling, the constant-builder case (Gasteiger et al., 2019a). The changing weights, rather than input retention alone, supply the additional feedback.

Use the maximum-row norm $\|X\|_{\infty,2} = \max_i \|x_i\|_2$. Fix $R \ge \|A\|_{\infty,2}$, $R > 0$, and let $B_R = \{X : \|X\|_{\infty,2} \le R\}$. Suppose

$$\|\mathcal{P}(X) - \mathcal{P}(Y)\|_\infty \le L_P \|X - Y\|_{\infty,2}, \qquad \kappa_{\text{f}} = RL_P, \qquad m = \beta - \alpha\kappa_{\text{f}}, \tag{1.5}$$

for $X, Y \in B_R$, where the matrix norm is the maximum absolute row sum. The dimensionless quantity κ_{f} measures how strongly changed weights can act on bounded representations; m is the certified contraction margin (Section 7.1). Any state-dependent mask, normalization or bandwidth belongs to the builder and its bound.

Because $P = P(X)$ is rebuilt from its input, the Jacobian of $X \mapsto P(X)X$ contains a feedback term $(\mathrm{D}P)[U]X$ that is sign-indefinite even when every kernel is positive and reversible. The frozen average cannot amplify: holding the realized weights fixed gives a row-stochastic map, non-expansive in the maximum-row norm, each of whose outputs is a convex combination of the anchors (Section 12.1). The response of the equilibrium to its input need not be nonnegative or non-expansive. Amplification comes from the feedback, not from the average.

1.5 Three walks and a toolbox to carry along

Three small walks recur as test cases; the numbers they produce are reproduced by the verification suite whose ledger is in Appendix B.

Worked example 1.6 (The two-community chain)**.** $P = \left(\begin{smallmatrix} 1-h & h \\ h & 1-h \end{smallmatrix}\right)$ with $h \le \frac{1}{2}$, so the walk is lazy and reversible with eigenvalues $\{1, 1-2h\}$ and content gap $\beta = 2h$ (Section 2.3).

Worked example 1.7 (The lopsided triangle)**.** Three tokens with $S = \left(\begin{smallmatrix} 0.50 & 0.30 & 0.20 \\ 0.30 & 0.45 & 0.25 \\ 0.20 & 0.25 & 0.55 \end{smallmatrix}\right)$, doubly stochastic, so π is uniform, with eigenvalues $\{1, 0.3366, 0.1634\}$; add a rotation, $P_s = S + sC$ with $C = \left(\begin{smallmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{smallmatrix}\right)$ the circulation on three tokens. For $s \in [0, 0.2]$, P_s is a walk with reversible part S, flow $J = sC$ and $\Phi = \sqrt{3}\, s$.

Worked example 1.8 (The self-or-sink chain)**.** Token 1 absorbs; every other token keeps weight $1 - \delta$ on itself and gives δ to token 1, with $\delta \in (0, 1]$. The walk is lower triangular, has no positive stationary distribution, and its powers converge to $\mathbf{1}e_1^\top$, the matrix whose every row is $e_1^\top$.

The conventions, and a toolbox of standard facts that the proofs use, are collected once, in Section A.1.

1.6 What holds in trained networks

Exact for every network of the stated class. The object itself (Section 1.1): every softmax head, at every layer and input, is a row-stochastic kernel $P(X)$ whose outputs are convex combinations of its values. The path-lottery form $\mathbb{E}[P^K]$ describes powers of one operator; a trained stack is an ordered product of different $P(X)$, with MLPs between.

The operator theory of Bruce (2026i) concerns a fixed operator, a fixed sequence of operators, and the lottery over layers that a residual stream defines. A trained transformer rebuilds its attention at every layer from the representation the previous layers produced, and interleaves MLPs and normalization. The bounds of Proposition 2.20(i)–(ii) concern realized products and so hold however the layers arose, and its clause (iii) holds along any trajectory of the simplified block $X_{\ell+1} = (1 - b_\ell)X_\ell + b_\ell P_\ell X_\ell$; nothing here models the full block. The doubly exponential rank collapse of attention-only networks (Dong et al., 2021) arises from re-mixing and is outside the operator class. When the operator is rebuilt from the representation it moves, $P = P(X)$, the static theory describes the operator at a fixed point of the loop; whether the loop has a unique globally attracting fixed point is a separate question, which Chapter 7 takes up.

Chapter 2

Depth is a distribution

Chapter 1 read depth, restarts and residual streams as lotteries over path lengths: an amplifier applies the walk a random number K of times, and its normalized form is $f(P) = \mathbb{E}[P^K]$ (Definition 1.3). This chapter asks what that random number decides. Chapter 0 gives the answer in words (Section 0.4.2): what survives many layers of averaging depends not on the number of layers but on how often the signal takes a short route. Read as an architecture's *path law*, the distribution of K answers three different questions through three of its functionals: its head decides whether a distinction survives, its mixing budget decides how fast it is lost, and its tail decides how far information travels.

The path law and the lottery of Chapter 1 are one random variable, the number of mixing steps that a signal takes, so the chapter reads the results about it as one theory, in two forms. The general form comes first: a family of path laws retains a mode exactly when it keeps a fixed share of its paths in a bounded window (Section 2.2). The collapse-ratio form comes second: on a lazy reversible operator the same condition is exactly what keeps the collapse ratio bounded, the collapse ratio being the factor by which amplification makes consensus louder than content (Section 2.3). The mixing budget then sets the rate of loss (Section 2.4); the condition carries over to causal operators, to every operator and to stacks of different layers (Section 2.5); normalization placement is a choice of budget (Section 2.6); the measurements follow (Section 2.7); and Section 2.8 states what no path law can do.

2.1 Architectures as path laws

Let $P \in \mathbb{R}^{N\times N}$ be row-stochastic. A *path law* is the distribution of a random variable K on $\mathbb{N}_0 \cup \{\infty\}$, with $p_k = \mathbb{P}(K = k)$ and generating function $f(z) = \mathbb{E}[z^K]$, where $z^\infty = 0$ for $|z| \le 1$ and $P^\infty = 0$; mass at ∞ is allowed so that substochastic designs are included. The filter

$$f(P) = \sum_{k\ge 0} p_k P^k = \mathbb{E}[P^K] \tag{2.1}$$

has a probabilistic reading: row i of $f(P)$ is the subprobability distribution of a random walk on P started at i and stopped at an independent random time K, with the mass at $K = \infty$ discarded. Hence $f(P)$ is stochastic and preserves every stationary law and absorbing state of P when $K < \infty$ almost surely. It acts on an eigenvector $Pv = \mu v$ by the scalar $f(\mu)$, with no normality assumption on P (Section A.3). Table 2.1 lists the path laws of twelve propagation and diffusion designs, and their derivations follow it. Three rows deserve comment. Input restart and a local residual both add the input back. The residual carries the evolving state, and its law

Table 2.1: **A dictionary of path laws.** Each design is a normalized nonnegative combination of powers of one operator P; the last column is the gain on a mode $\mu \in (0,1)$ as depth grows at fixed hyperparameters. G denotes a geometric law, $\mathbb{P}(G=k) = r(1-r)^k$. For continuous restart the bound holds at every time t. Named methods refer to their linear propagation stage.

Design	Filter or update	Path law K	Gain as $L \to \infty$
Powers (SGC; pure attention)	P^L	$K = L$	$\mu^L \to 0$
Lazy residual (residual GCN)	$((1-p)I + pP)^L$	Binomial(L,p)	$(1-p(1-\mu))^L \to 0$
Layer averaging (LightGCN)	$\frac{1}{L+1}\sum_{k=0}^{L} P^k$	Unif$\{0,\dots,L\}$	$\sim \frac{1}{(L+1)(1-\mu)} \to 0$
Atom plus averaging (S^2GC)	$aI + \frac{1-a}{L}\sum_{k=1}^{L} P^k$	$a\delta_0 + (1-a)$ Unif$\{1,\dots,L\}$	$\to a$
Input restart (APPNP, label spreading)	$h_{\ell+1} = rh_0 + (1-r)Ph_\ell$	truncated G	$\to \frac{r}{1-(1-r)\mu}$
Resolvent (PPNP, regularized Laplacian)	$r[I-(1-r)P]^{-1}$	G	$\frac{r}{1-(1-r)\mu}$
Diffusion, Euler steps (DGC)	$(I + \frac{t}{L}(P-I))^L$, $L \geq t$	Binomial$(L, t/L)$	$\to e^{-t(1-\mu)}$
Depth-scaled residual	$\left(\frac{L}{L+\sigma}I + \frac{\sigma}{L+\sigma}P\right)^L$	Binomial$(L, \frac{\sigma}{L+\sigma})$	$\to e^{-\sigma(1-\mu)}$
Heat kernel (GDC, frozen GRAND)	$e^{t(P-I)}$	Poisson(t)	$e^{-t(1-\mu)}$
Matérn kernel (normalized)	$[(1-q)(I-qP)^{-1}]^\nu$	NegBin(ν, q)	$\left(\frac{1-q}{1-q\mu}\right)^\nu$
Fractional diffusion	$e^{-t(I-P)^\alpha}$, $0<\alpha<1$	compound Poisson–Sibuya, $\mathbb{E}K = \infty$	$e^{-t(1-\mu)^\alpha}$
Continuous restart	$\dot{X} = \beta(P-I)X + \rho(X_0 - X)$	Poisson(βS), $S = \min(\text{Exp}(\rho), t)$	$\geq \frac{\rho}{\rho+\beta}$

Binomial(L,p) drifts to long paths; restart re-injects the original input, and its law converges to a geometric law with fixed mean $(1-r)/r$. Uniform averaging keeps the input at every depth, yet the input's weight $1/(L+1)$ vanishes; S^2GC keeps a fixed weight a on it (Zhu & Koniusz, 2021). Fractional diffusion $e^{-t(I-P)^\alpha}$, $0<\alpha<1$, is a compound Poisson law whose jumps have the Sibuya distribution: its zero-hop mass is e^{-t} and its mean path length is infinite. A deep network with learned feature maps reduces to a path law only on the channels where its feature action is scalar. The linear residual stack $X_\ell = X_{\ell-1} + A_\ell X_{\ell-1} V_\ell$ expands as a sum over subsets of layers, $X_L = \sum_{\ell_1<\cdots<\ell_k} A_{\ell_k}\cdots A_{\ell_1} X_0 V_{\ell_1}\cdots V_{\ell_k}$, and becomes (2.1) after normalization when $A_\ell = \beta_\ell P$ with $\beta_\ell \geq 0$ and the carried channel satisfies $V_\ell w = w$ (the binomial case is due to Chen et al., 2022). Random or learned feature maps are handled in Section 2.4, where the same calculus applies to expected energies, and measured attention stacks in Section 8.3, where it applies to products of kernels.

Each row of Table 2.1 refers to a linear propagation stage on a fixed operator. For an undirected graph with degree matrix D, $H = D^{-1/2}WD^{-1/2} = D^{1/2}PD^{-1/2}$ with $P = D^{-1}W$, so $f(H) = D^{1/2}f(P)D^{-1/2}$: random-walk statements refer to P, spectral ones to H, and similarity does not remove negative eigenvalues. A named method is covered only through the stage stated here; nothing is claimed about its learned feature maps or its accuracy on its own benchmarks. The derivations follow row by row, and the last three items treat related designs.

- *Powers, lazy residuals and averaging.* SGC applies P^L (Wu et al., 2019). Because I and P commute, $((1-p)I+pP)^L = \sum_k \binom{L}{k} p^k (1-p)^{L-k} P^k$, the binomial law of residual GCNs (Chen et al., 2022). LightGCN's uniform combination of layer embeddings (He et al., 2020) gives $f(\mu) = (1-\mu^{L+1})/[(L+1)(1-\mu)]$, and S^{2}GC's filter $aI + \frac{1-a}{L}\sum_{k=1}^{L} P^k$ (Zhu & Koniusz, 2021) keeps $f(\mu) \geq a$ for $\mu \in [0,1]$. Uniform averaging is also the consensus-normalized form of unweighted input injection: $h_\ell = Ph_{\ell-1} + h_0$ gives $h_\ell/(\ell+1) = \frac{1}{\ell+1}\sum_{k\leq\ell} P^k h_0$. Injection at every layer therefore does not by itself retain; the injected input must keep a fixed relative weight. At $\mu = 0.9$ and $a = 0.05$, averaging and S^{2}GC have gains 0.860 and 0.822 at depth 3 but 0.078 and 0.117 at depth 128: the floor is asymptotic, not a ranking at every depth.

- *Input restart and resolvents.* Iterating $h_{\ell+1} = rh_0 + (1-r)Ph_\ell$ gives $h_L = [r\sum_{k<L}(1-r)^k P^k + (1-r)^L P^L]h_0$, APPNP's truncated geometric law (Gasteiger et al., 2019a), and on a mode $f_L(\mu) = \frac{r}{1-a} + a^L(1-\frac{r}{1-a})$ with $a = (1-r)\mu$. For $\mu \in (0,1)$ this is at least its limit $r/(1-(1-r)\mu)$ at every depth, which gives the restart rule of Section 21.1. The limit is PPNP, $r[I-(1-r)P]^{-1}$, which equals the regularized Laplacian $(I+\lambda(I-P))^{-1}$ with $r = 1/(1+\lambda)$ (Smola & Kondor, 2003) and solves $\min_Z \frac{1}{2}\|Z-X\|^2 + \frac{\lambda}{2}\operatorname{tr}(Z^\top(I-H)Z)$ in symmetric coordinates. Label spreading (Zhou et al., 2003) and Correct & Smooth (Huang et al., 2021) iterate the same map, and linear input-injected equilibria (Bai et al., 2019; Gu et al., 2020) have the same resolvent on channels where their update reduces to this restart map. GCNII's initial residual (Chen et al., 2020a) reduces to restart when its feature maps are the identity.

- *Diffusion.* For a frozen operator, $\dot{X} = (P-I)X$ gives $e^{t(P-I)} = \sum_k e^{-t} t^k P^k/k!$, the Poisson law; this is uniformization (Norris, 1997), and it covers heat kernels, GDC's heat diffusion and frozen linear GRAND (Smola & Kondor, 2003; Gasteiger et al., 2019b; Chamberlain et al., 2021). Forward Euler with $L \geq t$ steps gives DGC's $(I + \frac{t}{L}(P-I))^L$, the law Binomial$(L, t/L)$ (Wang et al., 2021); for $L < t$ the identity coefficient is negative and the design leaves the class. The normalized depth-scaled residual $(I+\sigma P/L)^L/(1+\sigma/L)^L = (\frac{L}{L+\sigma}I + \frac{\sigma}{L+\sigma}P)^L$ has law Binomial$(L, \sigma/(L+\sigma))$ and converges in distribution to Poisson(σ). For $\mu \in (0,1)$, its gain approaches $e^{-\sigma(1-\mu)}$ from above, while DGC's Euler gain with $t = \sigma$ approaches that value from below.

- *Matérn kernels.* With $\mathcal{L} = I - H$, length scale κ, $c = 2\nu/\kappa^2$ and $q = 1/(1+c)$, the normalized graph Matérn covariance is

$$c^\nu(cI+\mathcal{L})^{-\nu} = \left[(1-q)(I-qH)^{-1}\right]^\nu = \sum_k \frac{\Gamma(k+\nu)}{\Gamma(\nu)k!}(1-q)^\nu q^k H^k,$$

  a negative binomial law with mean $\nu q/(1-q) = \kappa^2/2$ and zero-hop mass $(1-q)^\nu$ (Borovitskiy et al., 2021). The original construction uses the unnormalized Laplacian $\Delta = D - W$; choosing $b \geq \max_i D_{ii}$ and $P_b = I - \Delta/b$ (symmetric and stochastic) gives $c^\nu(cI+\Delta)^{-\nu} = [(1-q_b)(I-q_bP_b)^{-1}]^\nu$ with $q_b = b/(c+b)$, so that construction is also a path law, with a mean $\nu b/c$ that depends on the clock rate b.

- *Fractional diffusion.* For $0 < \alpha < 1$, $1-(1-z)^\alpha = \sum_{k\geq1} b_k z^k$ with $b_k = (-1)^{k+1}\binom{\alpha}{k} = \alpha\Gamma(k-\alpha)/[\Gamma(1-\alpha)\Gamma(k+1)] > 0$ and $\sum_k b_k = 1$: the Sibuya law, whose mean is infinite because the derivative $\alpha(1-z)^{\alpha-1}$ diverges at $z = 1$. A Poisson(t) number of Sibuya jumps has generating function $\exp\{t(g(z)-1)\} = e^{-t(1-z)^\alpha}$, with $g(z) = 1-(1-z)^\alpha$ the Sibuya generating function, so $e^{-t(I-H)^\alpha} = \mathbb{E}[H^K]$ with $\mathbb{P}(K=0) = e^{-t}$ and $\mathbb{E}K = \infty$. This is fractional diffusion of the normalized Laplacian $I-H$. Fractional graph models can instead use a fractional power of normalized adjacency (Maskey et al., 2023). At $t = 1$ and $\alpha = \frac{1}{2}$, its gain is 0.7289 at $\mu = 0.9$

against 0.9048 for heat, and 0.2938 at $\mu = -0.5$ against 0.2231: the heavy tail does not make it uniformly less attenuating.

- *Continuous restart.* For $\dot{X} = \beta(P - I)X + \rho(X_0 - X)$ with $X(0) = X_0$, variation of constants gives $X(t) = e^{-\rho t}e^{\beta t(P-I)}X_0 + \int_0^t \rho e^{-\rho u}e^{\beta u(P-I)}X_0\,du = \mathbb{E}[P^K]X_0$ with $K \mid S \sim \text{Poisson}(\beta S)$ and $S = \min\{\text{Exp}(\rho), t\}$, so $\mathbb{P}(K = 0) = \frac{\rho}{\rho+\beta} + \frac{\beta}{\rho+\beta}e^{-(\rho+\beta)t} \geq \frac{\rho}{\rho+\beta}$. CGNN's unweighted source injection (Xhonneux et al., 2020) gives a different equation, as do selective source terms.

- *Aggregation, halting and sampling.* A fixed convex combination of the same P^kX is a path law, but concatenation, feature-specific maps, adaptive or signed weights, as in JK-Net, SIGN and DAGNN (Xu et al., 2018; Frasca et al., 2020; Liu et al., 2020b), generally leave the class. Halting mechanisms define a scalar filter only when the output is a global nonnegative mixture of the states P^kX; sampling an iteration count inside an expected loss, as in recurrent-depth training (Geiping et al., 2025), is not the same operation as applying the expected operator. Finally, for a path law with $K < \infty$ almost surely, sampling K and a K-step walk from i and returning x_{X_K} is an unbiased estimator of $(f(P)x)_i$, with Hoeffding concentration and expected cost $\mathbb{E}K$ per sample, which is infinite for fractional diffusion even though every sampled path is finite.

- *Residual-stream mixing.* Hyper-connections mix several residual streams, and mHC constrains the mixers to be doubly stochastic (Xie et al., 2025). For exact doubly stochastic M_ℓ on n streams, $M_L \cdots M_1 - E = (M_L - E)\cdots(M_1 - E)$ with $E = \mathbf{1}\mathbf{1}^\top/n$, so the stream-consensus distance of the product is a diagnostic of the kind in Section 8.3. Positivity of every mixer does not force consensus: $M_\ell = (1 - \epsilon_\ell)I + \epsilon_\ell E$ with $\epsilon_\ell = (\ell + 1)^{-2}$ is strictly positive, yet $M_L \cdots M_1 = E + \frac{L+2}{2(L+1)}(I - E) \to \frac{1}{2}(I + E)$. For this form with $0 \leq \epsilon_\ell < 1$ at every layer, consensus requires $\sum_\ell \epsilon_\ell = \infty$, a divergent budget in the sense of Section 2.4; the infinite-flow property is the general counterpart (Touri & Nedić, 2011).

- *Gated recurrences.* For fixed gates $a_t \in [0, 1]$, $h_t = a_t h_{t-1} + (1 - a_t)x_t$ is a causal stochastic kernel on $(h_0, x_1, \ldots, x_T)$ with $P_{ts} = (1 - a_s)\prod_{j=s+1}^t a_j$ and an absorbing initial state, so stacks of such channelwise layers satisfy the survival identity of Section 8.3. The same product criterion governs memory along the sequence: for an input with $a_s < 1$ and positive later gates, the weight on x_s stays positive as $t \to \infty$ if and only if $\sum_{j>s}(1 - a_j) < \infty$. Selective state-space models have non-normalized maps and need their own reduction (Gu & Dao, 2024; Wang et al., 2025).

2.2 The head decides retention

A family of designs (one architecture at all depths, or all settings of a hyperparameter) retains a mode if its gain stays bounded away from zero across the family.

Theorem 2.1 (The head decides retention)**.** *Fix $\mu \in (0, 1)$ and a nonempty family of path laws $\{K_\theta\}$ with generating functions f_θ. Then*

$$\inf_\theta f_\theta(\mu) > 0 \quad \Longleftrightarrow \quad \exists\, m \in \mathbb{N}_0,\ c > 0 : \quad \mathbb{P}(K_\theta \leq m) \geq c \ \text{ for all } \theta.$$

Quantitatively, for every law and every $m \in \mathbb{N}_0$, with $q_m = \mathbb{P}(K \leq m)$,

$$q_m\,\mu^m \ \leq\ f(\mu) \ \leq\ q_m + (1 - q_m)\,\mu^{m+1}, \tag{2.2}$$

so $f(\mu) \geq \gamma$ forces $q_m \geq (\gamma - \mu^{m+1})/(1 - \mu^{m+1})$. The upper bound is attained by a law on $\{0, m+1\}$.

Proof. Split $\mathbb{E}[\mu^K]$ at m: paths of length at most m contribute between μ^m and 1 each, longer paths at most μ^{m+1}. This gives (2.2). If the gains are at least $\gamma > 0$, choose m with $\mu^{m+1} < \gamma$; the upper bound gives $q_m \geq (\gamma - \mu^{m+1})/(1 - \mu^{m+1}) > 0$ uniformly. Conversely $q_m \geq c$ gives $f \geq c\mu^m$. □

The sharpness of both bounds, a certificate and the equivalent forms of the criterion are in Section A.3. A numerical check finds the window bounds (2.2) exact to machine precision on 2,000 random path laws (largest violation 1.1×10^{-16}). Equivalently, a sequence of designs loses the mode, $f_n(\mu) \to 0$, exactly when $K_n \to \infty$ in probability. Along a single sequence this is the continuity theorem for generating functions (Feller, 1968); what Theorem 2.1 adds is the family criterion with constants, and the following consequences, each of which contradicts a common reading of depth. *The mean is irrelevant.* A bounded mean suffices, since $f(\mu) \geq \mu^{\mathbb{E}K}$ by Jensen's inequality, but it is not necessary. The law $\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{2L-1}$ has mean L and keeps gain above $\mu/2$ at every depth; the law $\frac{1}{L}\delta_1 + (1 - \frac{1}{L})\delta_{L+1}$ has the same mean and loses everything. On Cora the two retain 18.5% and 0.5% of the label energy at $L = 256$ (Figure 2.1c), and fractional diffusion retains even with an infinite mean. *An identity path is not necessary, and averaging is not enough.* The law $K \equiv 1$ retains μ with no zero-hop mass. Conversely, uniform averaging includes the input at every depth, yet its window mass $\mathbb{P}(K \leq m) = (m+1)/(L+1)$ vanishes: LightGCN-style averaging forgets every fixed mode, at rate $1/((L+1)(1-\mu))$. A fixed input weight a restores a floor a, which is why S^{2}GC and restart retain. *One mode is all modes.* A window gives $f_\theta(x) \geq c\, x^m$ for all $x \in [0, 1]$. It follows that retaining one interior positive mode is equivalent to retaining all of them, with mode-dependent floors. The operator must be held fixed: if $\mu_L = 1 - 1/L$ drifts with depth for $L \geq 2$, $K = L$ has gain $(1 - 1/L)^L \to e^{-1}$ without any window. The escape direction does not need a positive eigenvalue, reversibility or diagonalizability. The retention criterion requires a fixed positive interior eigenvalue, but no reversibility or diagonalizability.

Corollary 2.2 (Escape is universal; retention is positive)**.** (a) *If $K_n \to \infty$ in probability, then for every stochastic P, $f_n(P)x \to 0$ for every x in the span of the generalized eigenvectors of P with eigenvalues $|\lambda| < 1$. For irreducible aperiodic P with stationary law π, $f_n(P) \to \mathbf{1}\pi^\top$ for proper laws.* (b) *For $\mu \in (0, 1)$, $|f(-\mu)| \leq f(\mu)$ and $f(-\mu) \leq f(\mu)$, with equality in the latter if and only if every finite path length with positive mass is even.*

The proof is in Section A.3. Part (a) covers the non-normal, possibly defective kernels of causal attention (Sections 2.5 and 8.3). Part (b) is a rectification principle: a nonnegative filter passes a negative mode with the correct sign at most as well as its positive twin, and equally well only through even powers. For negative and complex modes the converse of (a) fails through cancellation: $f(z) = \frac{1}{3} + \frac{2}{3}z$ keeps all its mass on $\{0, 1\}$ yet vanishes at $z = -\frac{1}{2}$.

Some designs can be split into arbitrarily many identical stages: diffusion for time t is diffusion for time t/n repeated n times, and a residual ODE is the limit of ever finer residual stacks. Call a path law *refinable* if for every $n \geq 1$ it is the law of a sum of n independent and identically distributed $\mathbb{N}_0$-valued steps. For such designs the head collapses to a single number.

Proposition 2.3 (Refinable laws retain through their atom)**.** *A path law on $\mathbb{N}_0$ is refinable if and only if $f(z) = \exp\{-\lambda(1 - h(z))\}$ for some rate $\lambda \geq 0$ and some generating function h of a law on $\{1, 2, \dots\}$. Then $\mathbb{P}(K = 0) = e^{-\lambda}$ and*

$$e^{-\lambda} \;\leq\; f(\mu) \;\leq\; e^{-\lambda(1-\mu)} \qquad (\mu \in [0, 1]).$$

Consequently a family of refinable designs retains every positive mode if and only if $\sup_\theta \lambda_\theta < \infty$, if and only if $\inf_\theta \mathbb{P}(K_\theta = 0) > 0$, whatever the jump laws h_θ.

The proof is in Section A.3. The characterization is Feller's: infinitely divisible laws on $\mathbb{N}_0$ are compound Poisson (Feller, 1968, Chapter XII); the operator form is discrete subordination (Dungey, 2011; Bendikov & Saloff-Coste, 2012). Heat diffusion has $h(z) = z$ and $\lambda = t$; fractional diffusion has the Sibuya jumps $h(z) = 1 - (1 - z)^\alpha$ and $\lambda = t$; geometric and negative binomial laws have logarithmic-series jumps. Binomial laws and fixed $K = L \geq 1$ are not refinable. In continuous depth, then, retention is a clock rate: a design keeps a distinction if and only if its total rate of mixing events is bounded, and the jump law, heavy tails and infinite means included, is free to serve reach.

For proper path laws, retention compares a mode with the consensus, which is the right comparison when the consensus is itself a nuisance, such as item popularity in recommendation, which models store in the principal spectrum (Lin et al., 2025), or a class prior. A readout that removes the consensus before taking a class-score argmax cares instead about the relative gains of the non-consensus modes. Exponential tilting turns this into the same criterion.

Corollary 2.4 (Shape is retained by the tilted head)**.** *For $0 < \mu_1 < \mu_2 \leq 1$ and a law with $f(\mu_2) > 0$, let $\widetilde{K}$ have the tilted law $\mathbb{P}(\widetilde{K} = k) = p_k \mu_2^k / f(\mu_2)$. Then $f(\mu_1)/f(\mu_2) = \mathbb{E}[(\mu_1/\mu_2)^{\widetilde{K}}]$, so Theorem 2.1 applied to $\widetilde{K}$ decides whether relative structure survives. When $\mu_2 < 1$, uniform averaging keeps it at every depth: $f(\mu_1)/f(\mu_2) \geq 1 - \mu_2$.*

Uniform averaging therefore loses consensus-relative gain yet keeps positive non-consensus modal gains in bounded ratios. Section 2.7 reports its retained energy and label-propagation accuracy. The tail of the law governs a third quantity. On a graph, a node at distance d from node i receives weight $f(P)_{ij} \leq \mathbb{P}(K \geq d)$, because shorter walks cannot reach it.

Proposition 2.5 (Retention caps reach)**.** *If $f(\mu) \geq \gamma$ for some $\mu \in (0, 1)$, then for every $d \geq 1$*

$$\mathbb{P}(K \geq d) \;\leq\; \frac{1 - \gamma}{1 - \mu^d},$$

with equality for the law on $\{0, d\}$ whenever the right side is at most one.

The proofs of Corollary 2.4 and Proposition 2.5 are in Section A.3. For refinable laws the two functionals decouple: retention lives in the clock rate λ and reach in the jump law. Fractional diffusion keeps its atom e^{-t} while its tail decays only polynomially, $\mathbb{P}(K \geq d) \sim t\, d^{-\alpha}/\Gamma(1 - \alpha)$, against $(1 - r)^d$ for a geometric law with the same atom. Reach is necessary but not sufficient for delivering distant information through bottlenecks (Alon & Yahav, 2021; Topping et al., 2022); Proposition 2.5 states only what the path law permits. *When paths outrun memory.* For a proper path law and any stochastic P with stationary law π, write $d(k) = \max_i d_{\mathrm{TV}}(P^k(i, \cdot), \pi)$ and $t(\varepsilon) = \min\{k : d(k) \leq \varepsilon\}$. Convexity of total variation gives

$$\|f(P) - \mathbf{1}\pi^\top\|_\infty \;\leq\; 2\,\mathbb{E}\, d(K) \;\leq\; 2\,\mathbb{P}(K < t(\varepsilon)) + 2\varepsilon. \tag{2.3}$$

A distance from consensus larger than 2ε requires some paths shorter than the operator's mixing time. The operator sets the time scale and the law decides how much mass stays below it: Cora's propagation operator has relaxation time $1/(1 - \mu_2) = 276$ steps and filtered Chameleon's 172, while a random 4-regular graph of Cora's size relaxes in about 9. The same path law therefore preserves very different amounts on different graphs, and a depth budget is meaningful only relative to its operator.

Figure 2.1 draws the criterion on one graph. The head decides retention: a positive mode keeps a non-vanishing gain at every depth if and only if a fixed share of path mass stays inside a bounded window (Theorem 2.1). Five designs of the same depth have different path laws, and uniform

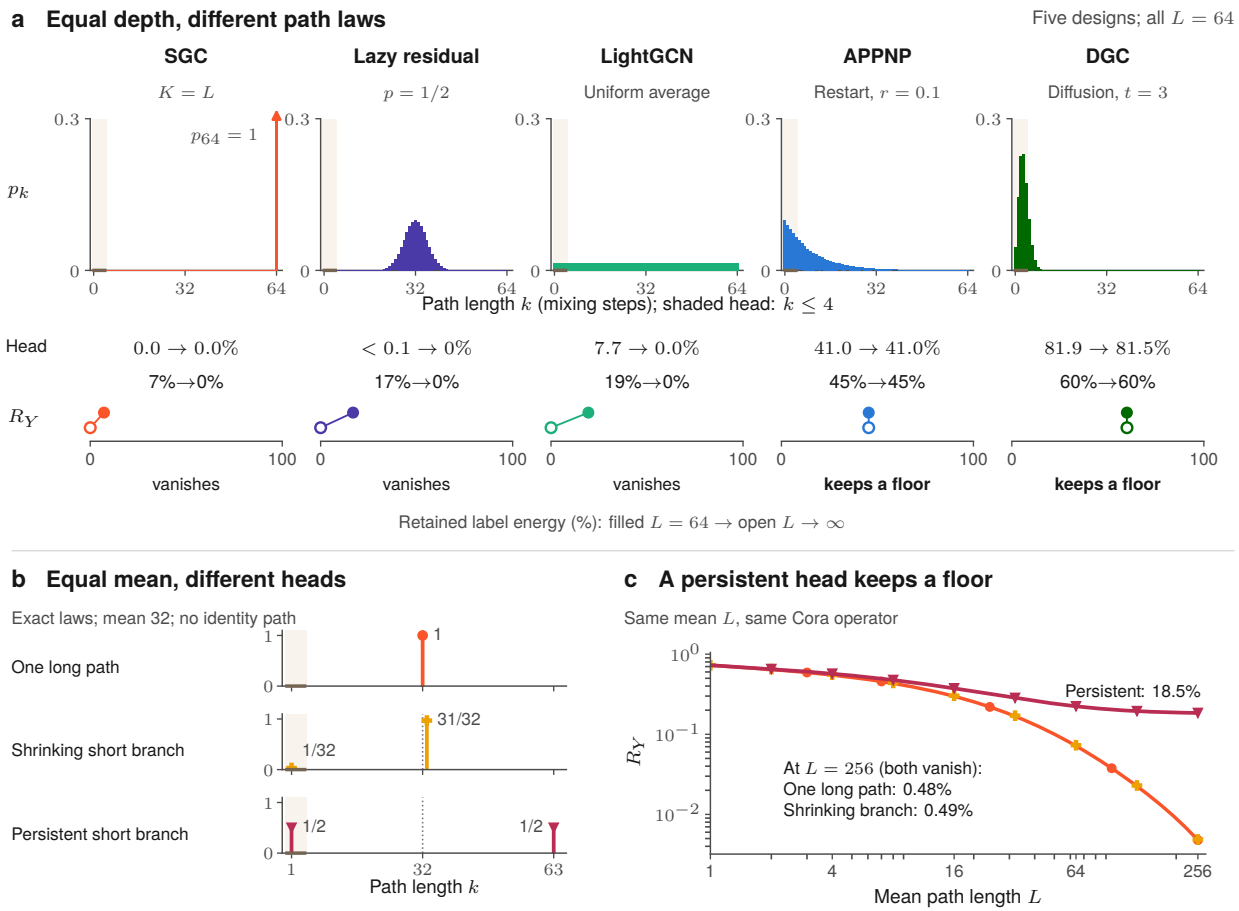


Figure 2.1: **A persistent short-path head keeps a retention floor.** (a) Exact illustrative PMFs $p_k = \mathbb{P}(K = k)$ at $L = 64$, all on the same 0–0.3 probability scale; the upward arrow at $k = 64$ marks the clipped SGC spike, whose true mass is $p_{64} = 1$. Shading and brown baselines mark $k \leq 4$. The head row gives $\mathbb{P}(K \leq 4)$ at $L = 64$ and its limit; the energy pairs give Cora's retained label energy R_Y (defined in Section 2.7) on 0–100% scales (filled: $L = 64$; open: $L \to \infty$). Uniform averaging's finite head vanishes; restart and fixed-time diffusion keep a fixed share. (b) Three equal-mean laws, δ_L, $L^{-1}\delta_1 + (1 - L^{-1})\delta_{L+1}$ and $\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{2L-1}$, shown at $L = 32$; none has an identity path. Dotted guides mark the mean. (c) Their Cora energy as the common mean grows; the shrinking branch is shown at depth octaves to expose its near coincidence with powers. Limits in (a) are analytic, using the saved spectrum for energy. One fixed Cora operator (2,485 nodes); deterministic spectral calculations, with no fitted uncertainty. The fixed-positive-mode criterion is Theorem 2.1; $k \leq 4$ is an example window. Records: `results/graphs/` (Bruce, 2026k).

averaging's finite head vanishes while restart and fixed-time diffusion keep a fixed share (panel a). The mean path length is irrelevant to retention: laws with the same mean, none of them with an identity path, can keep a floor or lose everything (panels b and c), and fractional diffusion retains every positive mode although its mean path length is infinite.

2.3 Collapse as a race between the lottery and a leak

The collapse-ratio form rests on one lemma about the expected survival $\mathbb{E}[c^K]$ of a lottery.

Lemma 2.6 (Survival)**.** *Let K be a nonnegative integer random variable and $c \in (0,1)$.*

(a) Truncation. *For every $m \geq 0$, $c^m \Pr[K \leq m] \leq \mathbb{E}[c^K] \leq \Pr[K \leq m] + c^m$.*

(b) Window. *If $\mathbb{E}[c^K] \geq s > 0$, then $\Pr[K \leq m] \geq s/2$ for every $m \geq \ln(2/s)/\ln(1/c)$.*

(c) Short-path mass. *For a family $\{K_\theta\}$, $\inf_\theta \mathbb{E}[c^{K_\theta}] > 0$ if and only if the family keeps* uniform short-path mass*: there are m and $\rho > 0$ with $\Pr[K_\theta \leq m] \geq \rho$ for all θ. The condition does not depend on $c \in (0,1)$.*

(d) Price. *For $0 < c_0 < c < 1$ and $q = \ln c / \ln c_0 \in (0,1)$, $\mathbb{E}[c^K] \leq (\mathbb{E}[c_0^K])^q$, with equality if and only if K is almost surely constant.*

Proof. (a) Split the expectation at m: the terms with $K \leq m$ are between c^m and 1 each, the rest at most c^m in total. (b) By the upper half of (a), $\Pr[K \leq m] \geq s - c^m \geq s/2$ once $c^m \leq s/2$. (c) If the family keeps short-path mass, the lower half of (a) gives $\mathbb{E}[c^{K_\theta}] \geq \rho c^m$. Conversely, if $\inf_\theta \mathbb{E}[c^{K_\theta}] = s > 0$, (b) gives the condition with $\rho = s/2$. (d) Write $c = c_0^q$, so $c^K = (c_0^K)^q$, and apply Jensen's inequality to the concave map $x \mapsto x^q$ (Section A.1, T6); strict concavity gives equality only when c_0^K, hence K, is almost surely constant. □

Uniform short-path mass is weaker than tightness in Prokhorov's sense: part of the mass may escape to infinity, provided a fixed share stays on paths of bounded length. Everything in this section and in Section 2.5 applies the lemma at $c = 1 - \text{leak}$, where the leak is the fraction of content that one averaging step destroys. Part (c) is the core of Theorem 2.8; part (d) says that suppressing an interior mode costs collapse, and that plain depth attains the lower bound at suppressions it can realize. The lemma and the characterization built on it (Theorem 2.8) are restated from Section 2.2, where the window bound (2.2) is sharper, attained by a lottery on $\{0, m+1\}$, and the price is sharpened to an exact two-mode frontier in Section 2.8.

Let P be lazy, reversible and irreducible, with eigenpairs (μ_k, v_k) as in Lemma 1.2. The *content modes* are the eigenvectors v_k with $k \geq 2$. A readout depends on some set of them; let μ_C be the largest eigenvalue in that set, the slowest to decay, and call $\beta = 1 - \mu_C$ the *content gap*, the fraction of the slowest readout mode's amplitude lost in one averaging step. An amplifier multiplies mode k by $g(\mu_k)$ and the consensus mode by $g(1)$, so the relevant number is the ratio.

Definition 2.7 (Collapse ratio)**.** $\kappa_{\text{col}}(g) = g(1)/g(\mu_C)$: how much louder the amplifier makes consensus than content.

Normalizing, $f(1) = 1$ and $f(\mu_C) = \mathbb{E}[\mu_C^K]$, so

$$\kappa_{\text{col}}(g) \;=\; \frac{1}{\mathbb{E}\big[(1-\beta)^K\big]}, \qquad \text{and} \qquad \|f(P) - \mathbf{1}\pi^\top\|_\pi \;=\; f(\mu_2) \;=\; \mathbb{E}\big[(1-\gamma)^K\big] \tag{2.4}$$

for the distance of the amplified operator from its rank-one limit, with $\gamma = 1 - \mu_2$ the spectral gap. The first is an identity; the second holds because $f(P) - \mathbf{1}\pi^\top$ is π-self-adjoint, kills $\mathbf{1}$ and multiplies v_k by $f(\mu_k)$ for $k \geq 2$, and f is nonnegative and increasing on the lazy spectrum $[0,1]$ (Section A.1, T1). The collapse ratio is the reciprocal of the expected survival of content over the network's paths. Table 1.1 evaluates it; in a machine-precision check on the two-community chain (Worked example 1.6) at $\beta = 0.05$, depth gives $\kappa_{\text{col}} = 2.27, 5.16, 26.6$ at $L = 16, 32, 64$; a restart at $\alpha = 0.9$ gives 1.45; a residual stream with total update $\sigma = 4$ gives $e^{0.2} = 1.221$; and a residual stack with

constant share $b = \frac{1}{2}$ gives 5.05 at $L = 64$, just below depth 32 at the same mean path length, as Proposition 2.10 requires. Which families of filters stay safe as they are strengthened? Let θ index strength (the depth, the mean $\alpha/(1-\alpha)$ of a restart, the total update σ); the interesting case is a family whose members suppress interior modes ever more deeply. The survival lemma gives the answer.

Theorem 2.8 (Short paths characterize bounded collapse)**.** *Fix a lazy reversible spectrum with content eigenvalue $\mu_C = 1 - \beta \in (0,1)$, and let $\{g_\theta\}$ be a family of amplifiers with lotteries K_θ and normalized forms f_θ. The following are equivalent:*

(i) bounded collapse: $\kappa^\star_{\mathrm{col}} := \sup_\theta \kappa_{\mathrm{col}}(g_\theta) < \infty$;

(ii) uniform short-path mass: $\Pr[K_\theta \le m] \ge \rho$ *for all* θ, *for some* m *and* $\rho > 0$;

(iii) a whole-spectrum floor: *for some* m *and* $\rho > 0$, $f_\theta(\mu) \ge \rho\,\mu^m$ *for all* $\mu \in [0,1]$ *and all* θ;

(iv) no deep suppression: $\inf_\theta f_\theta(\mu_0) > 0$ *at some, equivalently every,* $\mu_0 \in (0,1)$.

Moreover: (window) *if* $\kappa^\star_{\mathrm{col}} < \infty$, *then* $\Pr[K_\theta \le m^\star] \ge 1/(2\kappa^\star_{\mathrm{col}})$ *for every* θ, *with* $m^\star = \lceil \ln(2\kappa^\star_{\mathrm{col}})/\ln(1/\mu_C) \rceil$; (price) *for a single amplifier and an interior eigenvalue* $0 < \mu_0 < \mu_C$, *suppressing that mode by* $\Sigma = g(1)/g(\mu_0)$ *costs* $\kappa_{\mathrm{col}} \ge \Sigma^q$ *with* $q = \ln(1/\mu_C)/\ln(1/\mu_0) \in (0,1)$, *with equality if and only if the lottery is deterministic; and* (products) *for a composition* $f_L(P)\cdots f_1(P)$ *of hops of the same operator with per-hop anchors* $\hat a_\ell = f_\ell(0)$, *the collapse ratios* $\kappa_{\mathrm{col},L}$ *are bounded in* L *if and only if* $\sum_\ell (1 - \hat a_\ell) < \infty$.

Proof of (i) $\Longleftrightarrow$ (ii)*, the window, and the price.* By (2.4), (i) says $\inf_\theta \mathbb{E}[\mu_C^{K_\theta}] > 0$, which is (ii) by Lemma 2.6(c) at $c = \mu_C$. For the window, $f_\theta(\mu_C) = 1/\kappa_{\mathrm{col},\theta} \ge 1/\kappa^\star_{\mathrm{col}}$ and the upper truncation bound give $\Pr[K_\theta \le m^\star] \ge 1/\kappa^\star_{\mathrm{col}} - \mu_C^{m^\star} \ge 1/(2\kappa^\star_{\mathrm{col}})$. The price is Lemma 2.6(d) at $c = \mu_C$, $c_0 = \mu_0$. Clauses (iii), (iv) and (products) are proved in Section A.3. □

Clauses (i)–(iv), the window and the products clause restate, for the collapse ratio, Theorem 2.1 and Theorem 2.11, with their complements in Section A.3. An anchor $a \ge \rho$ is the simplest way to keep short-path mass (the case $m = 0$); it is sufficient, not necessary. A lottery with half its mass at $K = 1$ and half at $K = \theta$ has no anchor, and its collapse ratio stays below $2/\mu_C$ for every θ, as a synthetic check confirms. The window makes "short" relative: paths of up to $m^\star \approx \ln(2\kappa^\star_{\mathrm{col}})/\beta$ steps count, which on the content gap of a citation graph ($\beta \approx 0.026$) is about $38\ln(2\kappa^\star_{\mathrm{col}})$ steps. The price clause says the trade is real: no nonnegative filter both keeps collapse bounded and suppresses an interior mode arbitrarily, and plain depth pays the least at suppressions it can attain, although it collapses the most among lotteries with the same mean path length (Proposition 2.10). For the standard species the prices are explicit (Section 14.4). The products clause is the per-layer ledger: a stack of hops stays bounded exactly when its total anchor defect $\sum_\ell (1 - \hat a_\ell)$ is finite. A residual stack with a constant share b has defect b at every layer, anchor $(1-b)^L \to 0$ and collapse ratio $(1 - b\beta)^{-L} \to \infty$: it collapses, only more slowly than depth. A residual stream whose total update $\sum_\ell b_\ell$ stays below σ keeps short-path mass, since its path length satisfies $\Pr[K \le 2\sigma] \ge \frac{1}{2}$ by Markov's inequality, and hence a bounded collapse ratio, even if its anchor vanishes; with the shares $b_\ell = \sigma/(L+\sigma)$ it keeps an anchor $(1 + \sigma/L)^{-L} > e^{-\sigma}$ as well.

The equivalence is a statement about amplifiers with nonnegative coefficients, the filters that stacking and averaging produce, and not a law of nature. Filters with negative coefficients escape it. In synthetic checks at $\mu_C = 0.9$ (a scalar grid and one synthetic operator) (Bruce, 2026j), a brick-wall cutoff, a flat-top clip and a soft sigmoid all hold $\kappa_{\mathrm{col}} \approx 1$ while suppressing the interior deeply, which the price clause forbids in the nonnegative class; none admits a decomposition into nonnegative path weights, and the soft sigmoid's second derivative reaches -24.2 at sharpness

$\theta = 4$ ($-6{,}249$ at $\theta = 64$). A machine-precision check in the verification suite (Appendix B) gives a minimal instance at $\mu_C = 0.98$: a polynomial that equals 1 at the consensus and content modes, 0.01 at eight interior modes and 0 at the origin has $\kappa_{\text{col}} = 1$ exactly and suppresses the mode at 0.5 a hundredfold, which the price clause would charge $\kappa_{\text{col}} \geq 1.144$; it has five negative coefficients. The escape is paid in coefficient variation $\sum_k |c_k|$, which bounds $\|f(P)\|_\infty$ from above: suppressing a whole band below the content mode to ϵ_N costs a variation that grows at least polynomially in $1/\epsilon_N$ at any degree, and at a fixed degree one mode has an exact price. Write $f(z) = \sum_k c_k z^k$ with $f(1) = 1$ and let $n_- = \sum_k (-c_k)_+$ be its negative coefficient mass, so that $\sum_k |c_k| = 1 + 2n_-$. Since $\|P^k\|_\infty = 1$ for every row-stochastic P, $\|f(P)\|_\infty \leq \sum_k |c_k|$: coefficient variation bounds the filtered operator from above; it does not measure it.

Proposition 2.9 (The price of escape)**.** (i) One mode at a fixed degree. *Fix $0 \leq \mu_0 < 1$, a degree $L \geq 1$ and a target ϱ. Among polynomials of degree at most L with $f(1) = 1$ and $f(\mu_0) = \varrho$, the least negative mass is* $\operatorname{dist}(\varrho, [\mu_0^L, 1])/(1 - \mu_0^L)$*, attained using only the powers 0 and L. Exact suppression, $\varrho = 0$, costs $\mu_0^L/(1 - \mu_0^L)$.*

(ii) A band at any degree. *Let $0 < \mu_N < \mu_C < 1$, and require $|f| \leq \epsilon_N$ on $[0, \mu_N]$ and $|f - 1| \leq \epsilon_C < 1$ on $[\mu_C, 1]$. Choose $0 < a_0 \leq \mu_N$ and $\mu_C + a_0 < R < 1$, and put $z = 2\mu_C/a_0 - 1$, $\Lambda = z + \sqrt{z^2 - 1}$, $r = \mu_C/(R - a_0)$ and $\gamma = \ln(1/r)/\ln \Lambda > 0$. If $\epsilon_N \leq (1 - \epsilon_C)/(2\Lambda)$, every such polynomial, of any degree, has*

$$\sum_k |c_k| \;\geq\; \frac{(1 - \epsilon_C)(R - \mu_C)}{2R} \Big(\frac{1 - \epsilon_C}{2\epsilon_N} \Big)^{\gamma}.$$

(iii) The degree-free price grows polynomially. *For $\epsilon_N \leq \epsilon_C$ small enough, the least coefficient variation over all degrees lies between $C_- \epsilon_N^{-\gamma}$, with $C_- = (1 - \epsilon_C)(R - \mu_C)[(1 - \epsilon_C)/2]^\gamma/(2R)$, and $3\epsilon_N^{-K}$, with $K = 2 \ln 3/(\mu_C - \mu_N)^2$; the upper bound is achieved at degree $\lceil 2 \ln(1/\epsilon_N)/(\mu_C - \mu_N)^2 \rceil$.*

The proof is in Section A.3. The two exponents need not coincide, and the sharp exponent and constants of the degree-free price are not known. In synthetic checks, part (i) agrees with a linear program to 10^{-13}, and the inequality at the heart of (ii) holds with no violation on 2,563 (polynomial, m) pairs from 233 random polynomials of degree up to 39, with m the number of interpolation nodes of the proof. By Proposition 1.4, such filters are not averaging filters: they buy their escape by leaving the class. Such filters can help where the relevant modes sit at low or negative eigenvalues, as on the tested heterophilic graphs. At matched budgets, GPR-GNN's signed filters beat APPNP's monotone nonnegative ones on real data (Bruce, 2026j), with mean test-accuracy gains over 10 seeds of 24.3 and 28.2 percentage points on Texas and Wisconsin and 3.9 and 1.9 on Squirrel and Actor. "Bounded collapse needs short-path mass" is a theorem about a class, and the class boundary is part of the theorem.

Depth, restarts and residual streams amplify the same modes of the same operator, and to first order they are interchangeable: at equal mean path length $\bar{k}$ each gives $\ln \kappa_{\text{col}} \approx \bar{k}\beta$. They are not interchangeable at second order, and the lottery's dispersion decides how.

Proposition 2.10 (Dispersion law)**.** *Let K have mean $\bar{k}$ and a finite third moment. As $\beta \to 0$,*

$$\ln \kappa_{\text{col}} \;=\; \bar{k}\beta \;+\; \tfrac{1}{2}(\bar{k} - \operatorname{Var} K)\, \beta^2 \;+\; O(\beta^3).$$

Exactly: the Poisson lottery has $\ln \kappa_{\text{col}} = \bar{k}\beta$ for every β; the geometric lottery with mean $\bar{k}$ has $\ln \kappa_{\text{col}} = \ln(1 + \bar{k}\beta)$; depth has $\ln \kappa_{\text{col}} = -\bar{k} \ln(1 - \beta)$. At a fixed mean, depth collapses the most: $\kappa_{\text{col}} \leq (1 - \beta)^{-\bar{k}}$ for every lottery with $\mathbb{E}K = \bar{k}$ and every $\beta \in (0, 1)$.

Proof. With $s = -\ln(1-\beta) = \beta + \beta^2/2 + O(\beta^3)$, $\ln \kappa_{\mathrm{col}} = -\ln \mathbb{E}[e^{-sK}] = s\bar{k} - \frac{1}{2}s^2 \operatorname{Var} K + O(s^3)$ by the cumulant expansion (Section A.1, T16); substitute. The exact forms are direct evaluations. The last claim is Jensen's inequality for the convex map $x \mapsto (1-\beta)^x$: $\mathbb{E}[(1-\beta)^K] \geq (1-\beta)^{\bar{k}}$. □

At second order, under-dispersed lotteries (depth and constant-share stacks, $\operatorname{Var} K < \mathbb{E}K$) over-deliver collapse relative to the linear rate; over-dispersed ones (restarts, $\operatorname{Var} K = \bar{k} + \bar{k}^2$) under-deliver; the Poisson lottery sits exactly on the line. Trading one mechanism for another at matched first-order strength therefore carries a leading correction whose sign is fixed by dispersion alone. For depth, restarts and Poisson streams, the ordering is exact, with zero sign violations over 600,000 sampled $(\beta, \bar{k})$ pairs in a synthetic check (Bruce, 2026j). A tilted version of the same number returns in Section 3.5. The tails differ too: the Poisson lottery starves long paths factorially, $\Pr[K \geq m] \leq e^{-\sigma}(e\sigma/m)^m$ for $m \geq \sigma$, where a geometric lottery starves them only exponentially (Section A.3).

2.4 The mixing budget sets the rate

Most deep designs are stacks: layer ℓ contributes an independent number of steps K_ℓ, and the total path length is $K = K_1 + \cdots + K_L$. The gain then factorizes as $f(\mu) = \prod_\ell f_\ell(\mu)$. In this section, every layer is assumed to have positive probability of taking finitely many steps. A layer *mixes* if it takes at least one step. The expected number of mixing layers is the design's *mixing budget*,

$$S_L \;=\; \sum_{\ell=1}^{L} \mathbb{P}(K_\ell \geq 1). \tag{2.5}$$

Theorem 2.11 (The mixing budget sets the rate)**.** *Let $K = K_1 + \cdots + K_L$ with independent layer laws, which may depend on L, and let $\rho_L = \min_{\ell \leq L} f_\ell(\mu)$. For every $\mu \in (0,1)$,*

$$(1-\mu)\, S_L \;\leq\; -\log f(\mu) \;\leq\; S_L/\rho_L. \tag{2.6}$$

Proof. Every $k \geq 1$ has $1-\mu \leq 1-\mu^k \leq 1$, so $(1-\mu)\mathbb{P}(K_\ell \geq 1) \leq 1 - f_\ell(\mu) \leq \mathbb{P}(K_\ell \geq 1)$. Combine with $1 - x \leq -\log x \leq (1-x)/x$ on $(0,1]$ and sum over layers. □

Whenever the per-layer gains stay bounded below, the gain is $\exp(-\Theta(S_L))$ and depth has four regimes. The condition holds automatically for Bernoulli layers, where $\rho_L \geq \mu$, and for the energy laws below, where $\rho_L \geq \mu^2$.

Corollary 2.12 (Phase diagram of depth)**.** *With ρ_L bounded below, a design keeps a floor if and only if S_L is bounded; it forgets polynomially when $S_L \asymp \log L$, stretched-exponentially when $S_L \asymp L^\gamma$ with $0 < \gamma < 1$, and exponentially when $S_L \asymp L$. For Bernoulli layers with $\mathbb{P}(K_\ell = 1) = c/\ell$ the critical case is exact: $f(\mu) = \Gamma(L+1-a)/[\Gamma(1-a)\Gamma(L+1)] \sim L^{-a}/\Gamma(1-a)$ with $a = c(1-\mu)$.*

The proof is in Section A.3, and Bruce (2026k, Figure 2a) draws the four regimes. Two different experiments hide behind "making a network deeper". Extending one fixed sequence of layers keeps a floor if and only if $\sum_\ell \mathbb{P}(K_\ell \geq 1) < \infty$, a Borel–Cantelli condition; this is how a layer-index schedule such as $1/\ell$ or $1/\sqrt{\ell}$ behaves. Rebuilding every layer for each total depth, as depth-scaled recipes do, keeps a floor if and only if $\sup_L L\, p_L < \infty$ for a per-layer mixing probability p_L. DGC's fixed terminal time is the latter case (Wang et al., 2021): more Euler steps refine a fixed budget rather than lengthen the typical path, which is why DGC can take arbitrarily many steps. The escape direction also survives nonlinearity. For residual GCNs with augmented normalized

adjacency H and updates $X_\ell = (1-p_\ell)X_{\ell-1} + p_\ell \,\mathrm{ReLU}(HX_{\ell-1}W_\ell)$ with $\|W_\ell\|_2 \le 1$, the contraction lemmas of Oono & Suzuki (2020) and convexity of the distance $d_{\mathcal{M}}$ to their invariant subspace give $d_{\mathcal{M}}(X_L) \le \prod_\ell (1 - p_\ell(1-\lambda))\, d_{\mathcal{M}}(X_0) \le e^{-(1-\lambda)S_L} d_{\mathcal{M}}(X_0)$, where $\lambda < 1$ is their contraction constant: a divergent budget forces collapse with learned weights too (Section A.3). Retention does not transfer, because the bound is one-sided.

Residual branches with learned or random feature maps do not act as scalars, but their expected energies follow a path law. Consider $X_\ell = X_{\ell-1} + \alpha_\ell P X_{\ell-1} V_\ell$ with a fixed operator P, deterministic strengths $\alpha_\ell \ge 0$, and feature maps $V_\ell = m_\ell I + Z_\ell$ whose random part is centered and isotropic given the past: $\mathbb{E}[Z_\ell \mid \mathcal{F}_{\ell-1}] = 0$ and $\mathbb{E}[Z_\ell Z_\ell^\top \mid \mathcal{F}_{\ell-1}] = \sigma_\ell^2 I$. The *drift* $m_\ell \ge 0$ is the coherent, same-sign part of the branch. A mode input $X_0 = v w_0^\top$ with $Pv = \mu v$ stays of the form $v w_\ell^\top$; its expected amplification is $G_L(\mu) = \mathbb{E}\|w_L\|^2/\|w_0\|^2$, and its energy relative to consensus is $R_L(\mu) = G_L(\mu)/G_L(1)$.

Theorem 2.13 (Energy law with drift)**.** *In this model* $G_L(\mu) = \prod_{\ell=1}^L g_\ell(\mu)$ *with*

$$g_\ell(z) = 1 + 2\alpha_\ell m_\ell\, z + \alpha_\ell^2(m_\ell^2 + \sigma_\ell^2)\, z^2.$$

For $\mu \in [0,1]$, $R_L(\mu) = \mathbb{E}[\mu^{B_1+\cdots+B_L}]$ *with independent* $B_\ell \in \{0,1,2\}$, $\mathbb{P}(B_\ell = 1) = 2\alpha_\ell m_\ell / g_\ell(1)$ *and* $\mathbb{P}(B_\ell = 2) = \alpha_\ell^2(m_\ell^2 + \sigma_\ell^2)/g_\ell(1)$. *For a depth family with common* α_L, m_L, σ_L *in all* L *layers, relative energy at every mode* $\mu \in (0,1)$ *keeps a floor if and only if*

$$\sup_L \; L\,(\alpha_L m_L + \alpha_L^2 \sigma_L^2) < \infty. \tag{2.7}$$

The proof is in Section A.3. The squared norm $\|w\|^2$ is a product of two copies of the propagated signal. Centered noise moves both copies at once, which is why a noisy layer takes either zero or two steps, while drift moves one copy at a time. The budget is therefore $\sum_\ell (2\alpha_\ell m_\ell + \alpha_\ell^2\sigma_\ell^2)$ to leading order, and (2.7) is a coherence principle for residual scaling: *the depth exponent is set by the same-sign part of the branch.* Without drift, $\alpha_L = c/\sqrt{L}$ keeps the floor $e^{-c^2\sigma^2(1-\mu^2)}$; with any fixed drift $m > 0$ it forgets as $\exp(-2cm(1-\mu)\sqrt{L})$, and a floor requires $\alpha_L = O(1/L)$ (Figure 2.2a). The rule reconciles the two established scalings. Independent centered weights single out $1/\sqrt{L}$ and correlated ones $1/L$ (Marion et al., 2025; Hayou et al., 2021); depth-μP uses $1/\sqrt{L}$ (Yang et al., 2024a; Bordelon et al., 2024) and CompleteP uses $1/L$ (Dey et al., 2025). In the energy law the distinction is the drift: $1/\sqrt{L}$ is admissible exactly while $m_L = O(L^{-1/2})$. Layer-index schedules follow the same calculus with sums in place of L: $\alpha_\ell = c/\sqrt{\ell}$ without drift gives the budget $c^2\sigma^2 \log L$ and the exact relative energy

$$R_L(\mu) = \frac{\Gamma(L+1+c^2\sigma^2\mu^2)\,\Gamma(1+c^2\sigma^2)}{\Gamma(1+c^2\sigma^2\mu^2)\,\Gamma(L+1+c^2\sigma^2)} \asymp L^{-c^2\sigma^2(1-\mu^2)};$$

with $c = \sigma = 1$ and $\mu = 0.9$, relative energy at $L = 10^4$ is 0.186 under this schedule and 0.827 under $\alpha_L = 1/\sqrt{L}$. A slowly declining curve is not a floor. Bounded anisotropy gives two-sided energy bounds and preserves the sufficient condition for a floor. The converse is established when $\sigma_+^2\mu^2 < \sigma_-^2$, where σ_-^2 and σ_+^2 bound the noise covariance below and above (Section A.3).

Looped transformers reuse one block many times, which breaks the independence behind Theorem 2.13. Wang et al. (2026) argue that weight sharing correlates the residual updates across iterations, so that a block visited N times in each of M distinct positions needs the stronger scale $1/(N\sqrt{M})$ in place of $1/\sqrt{MN}$; Li et al. (2026) also derive a stronger scaling, from the way one shared parameter update is accumulated over the visits and read back by them during training. In the forward pass, weight sharing alone does not force the stronger scale, and the following exact

model says what does. In each of M blocks apply $X \leftarrow X + \varepsilon PXV_j$ exactly N times with the same map

$$V_j = \sigma s_j\,(\lambda I + \beta J), \qquad \lambda \in [0,1],\ \beta = \sqrt{1-\lambda^2}, \tag{2.8}$$

where the signs $s_j = \pm 1$ are independent and J is a skew-orthogonal map on features ($J^\top = -J$, $JJ^\top = I$), unrelated to the routing part of (1.2). Every choice of λ gives the same moments, $\mathbb{E}V_j = 0$ and $\mathbb{E}V_jV_j^\top = \sigma^2 I$; λ is the symmetric fraction of the shared map.

Proposition 2.14 (Tied loops: the symmetric part decides)**.** *Let $u = \varepsilon\sigma$ and $G_N(u) = \frac{1}{2}\big[(1+2\lambda u+u^2)^N + (1-2\lambda u+u^2)^N\big]$. Then*

$$R_{M,N}(\mu) = \Big[\frac{G_N(u\mu)}{G_N(u)}\Big]^M \ \geq\ \exp\Big\{-MN\,\varepsilon^2\sigma^2(1-\mu^2)\big[1+2(N-1)\lambda^2\big]\Big\}.$$

Hence the tied scale $\varepsilon = c/(N\sqrt{M})$ keeps $R_{M,N} \geq e^{-2c^2\sigma^2(1-\mu^2)}$ for every λ, M and N, and the total-depth scale $\varepsilon = c/\sqrt{MN}$ keeps $R_{M,N} \geq e^{-c^2\sigma^2(1-\mu^2)(1+2N\lambda^2)}$. Conversely, for a fixed mode $\mu \in [0,1)$ and fixed $c\sigma > 0$, under $\varepsilon = c/\sqrt{MN}$ the floor is lost as soon as $N\lambda^2 \to \infty$, uniformly in M; for fixed $\lambda \in (0,1]$ and fixed M, $\log R_{M,N}(\mu) = -2c\sigma\lambda(1-\mu)\sqrt{MN} + O(1)$.

The proof is in Section A.3. To leading order in ε, and as a rigorous bound, each visit to a tied block carries effective energy $\varepsilon^2\sigma^2[1+2(N-1)\lambda^2]$: the symmetric part of the map adds coherently with the other $N-1$ visits, the skew part does not. At the total-depth scale, a floor is kept when $N\lambda^2$ stays bounded and lost when $N\lambda^2 \to \infty$. The tied scale $1/(N\sqrt{M})$ keeps a floor for every λ, while a purely skew map (the rotational extreme, $\lambda = 0$) needs only $1/\sqrt{MN}$. Skew-symmetric recurrent and graph designs are this extreme (Chang et al., 2019; Gravina et al., 2023). Figure 2.2b shows the threshold: with $M = 4$ and $\mu = \frac{1}{2}$ under $\varepsilon = 1/\sqrt{MN}$, relative energy at $N = 10^4$ is 0.135 for $\lambda = N^{-1/2}$ and 2×10^{-5} for a fixed $\lambda = 0.05$. In this model weight sharing matters only through the alignment it creates, and alignment sets the exponent, as it does for width exponents (Everett et al., 2024).

2.5 Causal operators, every operator, and residual stacks

Deep autoregressive attention is the case the π-geometry cannot handle. We call a row-stochastic matrix *strictly causal* if it is lower triangular (the diagonal is allowed). Its first row has a single admissible entry, so $P_{11} = 1$ and token 1 is absorbing. Write

$$P = \begin{pmatrix} 1 & 0 \\ r & Q \end{pmatrix}, \qquad Q\mathbf{1} + r = \mathbf{1},$$

with $r_i = P_{i1}$ the one-step mass sent to the first token and Q the transient block, and let $\delta = \min_{i\geq 2} P_{i1}$ be the *sink weight*. Softmax attention puts positive mass on every admissible position, so $\delta > 0$. Let τ_i be the time the walk started at token i takes to reach token 1, and let the lottery K be independent of the walk.

Proposition 2.15 (The stationary geometry degenerates)**.** *Let P be strictly causal with $P_{i1} > 0$ for all $i \geq 2$.* (a) *$\pi = e_1$ is the unique stationary distribution, so $\|x\|_\pi = |x_1|$ is only a seminorm.* (b) *For the teleported regularization $P_\varepsilon = (1-\varepsilon)P + \varepsilon\mathbf{1}\mathbf{1}^\top/N$, the stationary distribution has $\pi_\varepsilon(1) = 1 - O(\varepsilon)$ and condition number $\sqrt{\pi_{\max}/\pi_{\min}} = \Theta(\varepsilon^{-1/2})$.*

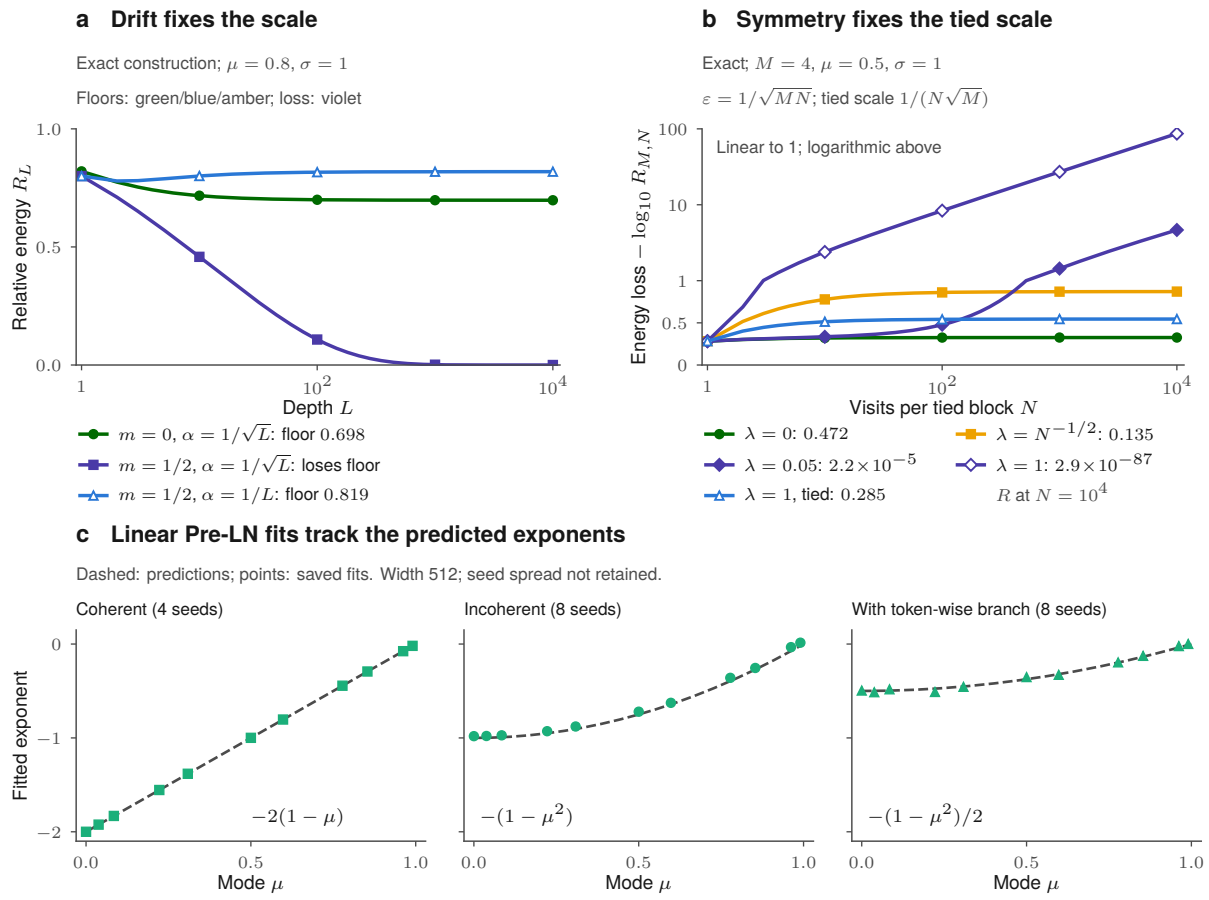


Figure 2.2: **Drift and a tied map's symmetric part decide the scale that retains energy.** (a) Exact illustrative construction from Theorem 2.13, with $\mu = 0.8$ and $\sigma = 1$. The total-depth scale $\alpha_L = 1/\sqrt{L}$ has the positive limit $e^{-0.36} \approx 0.698$ without drift, but loses the floor when $m = \frac{1}{2}$; $\alpha_L = 1/L$ restores the limit $e^{-0.2} \approx 0.819$. The linear ordinate separates the two positive limits. (b) Exact illustrative construction from Proposition 2.14, with $M = 4$, $\mu = \frac{1}{2}$, $\sigma = 1$ and $\varepsilon = 1/\sqrt{MN}$, except for the open triangles at the tied scale $1/(N\sqrt{M})$. The ordinate is the number of decades of relative energy lost, $-\log_{10} R_{M,N}$, drawn linearly from 0 to 1 and logarithmically above 1; it shows all five curves through $N = 10^4$, including the aligned map's 86.53-decade loss. The keys give the corresponding endpoint energies: $\lambda = N^{-1/2}$ retains 0.135, while fixed $\lambda = 0.05$ retains 2.2×10^{-5}. Green, blue and amber distinguish floor arms; violet denotes loss to zero in (a,b). (c) Dashed mean-field predictions against all eleven saved mode fits for each linear Pre-LN stack on the 32-token lazy ring, with per-token RMS normalization and width 512. The coherent case uses four seeds; the incoherent and token-wise-branch cases use eight each. Slopes are fitted to log-space averaged energies on $L \in \{128, 256, 512, 1024\}$; only the aggregates are retained, so seed spread and uncertainty intervals are unavailable. The three facets share the mode and exponent scales; mint identifies the critical polynomial regime. Predictions are $-2(1-\mu)$, $-(1-\mu^2)$ and $-(1-\mu^2)/2$ ($s = \frac{1}{2}$ in the last case), from Proposition 2.21 and its proof (Section A.3); measurements are the fits in `results/linear_stacks.json` (Bruce, 2026k).

The proof is in Section A.3.

Every π-metric statement therefore loses its meaning at the causal end, and any pipeline that computes a positive π for causal attention is regularizing it, with constants that blow up as the regularization is removed. The collapse law survives the loss of the metric by changing norms. We use $\|A\|_\infty = \max_i \sum_j |A_{ij}|$, the largest row-wise ℓ_1 size, which for differences of row-stochastic matrices is twice the worst row's total variation; the largest possible distance is 2.

Theorem 2.16 (Sink-collapse law)**.** *For every strictly causal row-stochastic P and every normalized amplifier f with lottery K and anchor a:*

(i) Identity. $\|f(P) - \mathbf{1}e_1^\top\|_\infty = 2\max_{i\geq 2} \Pr[K < \tau_i]$*: the distance to full collapse onto the first token is twice the worst token's probability that the lottery stops before the walk is absorbed.*

(ii) Envelope. $\|f(P) - \mathbf{1}e_1^\top\|_\infty \leq 2\,\mathbb{E}[(1-\delta)^K]$*, with equality for the self-or-sink chain (Worked example 1.8). For depth L the distance is at most $2(1-\delta)^L$: for $\delta > 0$, half-collapse within $\lceil \ln 2/\delta \rceil$ applications.*

(iii) Anchor floor. $\|f(P) - \mathbf{1}e_1^\top\|_\infty \geq 2a$*: an anchored lottery never collapses completely. For $\delta \in (0,1)$, the largest distance over strictly causal operators with sink weight at least δ is exactly $2\mathbb{E}[(1-\delta)^K]$, so a family of amplifiers keeps this class-extremal distance bounded below if and only if its lotteries keep uniform short-path mass.*

(iv) Rates. $\|Q^L\|_\infty \leq (1-\delta)^L$ *is nonincreasing in L, and the asymptotic rate is the spectral radius of Q, which equals* $\max_{i\geq 2} P_{ii}$.

The proof is in Section A.3. It rests on one sign pattern: each row of $f(P) - \mathbf{1}e_1^\top$ is a nonpositive entry on the first token followed by nonnegative entries, so their absolute values add without cancellation. Part (i) and the envelope in (ii) are the single-operator case of the survival identity of Section 8.3, which also covers stacks of different causal kernels with a residual share for every token; the anchor floor, the class extreme and the rates are added by Bruce (2026i). In a machine-precision check on the self-or-sink chain with $\delta = 0.1$, depth leaves distance 0.957 at $L = 7$ and 0.197 at $L = 22$ (half-collapse at $L \approx 6.6$), while a residual stack with limiting total update $\sigma = 2$ stays at $2(1 - \sigma\delta/(L+\sigma))^L$, which is 1.69 at $L = 10$ and tends to $2e^{-\sigma\delta} = 1.637$, above its anchor floor $2e^{-\sigma} = 0.271$. Three remarks fix the scope.

1. The equivalence in (iii) is class-extremal and cannot be made pointwise: at a fixed operator, a lottery with short-path mass but no anchor can collapse completely (in a machine-precision check, three tokens with $Q^2 = 0$ and $K \equiv 2$ give distance 0). Only the floor $2a$ holds at every operator.

2. The monotone decay in (iv) is specific to the ℓ_∞ reading: the transient block can grow in other norms (in a machine-precision check, Q with rows $(0.9, 0, 0), (0.9, 0, 0), (0, 0.9, 0)$ has $\|Q\|_1 = 1.8$).

3. The identity prices everything through a race between the lottery and absorption: a token with no effective path to the first token at the lottery's typical lengths has $\Pr[K < \tau_i] \approx 1$ and does not collapse.

Sinks. Token 1 is absorbing for every strictly causal operator. For a fixed such operator with $\delta > 0$, repeated application converges to it. It is also where trained language models place attention sinks (Xiao et al., 2024; Sun et al., 2024). The theorem does not explain why they do, or what the sink mass does for them; the functional accounts are that the pattern slows the mixing of content tokens (Barbero et al., 2025a), with the first token acting as a key-side bias whose value contribution is small (Gu et al., 2025). In the scalar-share mixing model, the residual stream's anchor keeps a composite from following a sink-heavy head to full collapse, as quantified in Proposition 2.20.

The reversible law is exact in the π-norm and the causal law in the ℓ_∞ norm. A single norm, and a single leak, cover every row-stochastic operator.

Definition 2.17 (Uniform leak)**.** For row-stochastic P on $N \geq 2$ tokens,

$$\lambda(P) = \min_{i \neq i'} \sum_j \min(P_{ij}, P_{i'j}) \in [0, 1],$$

the smallest overlap between two rows; equivalently $1 - \lambda(P) = \max_{i \neq i'} d_{\mathrm{TV}}(P_{i\cdot}, P_{i'\cdot})$, Dobrushin's coefficient (Dobrushin, 1956).

On the two-community chain, $\lambda = \gamma = \beta = 2h$. For every strictly causal operator $\lambda = \delta$ exactly: row 1 is $e_1^\top$, so its overlap with row i is P_{i1}, and any two rows share at least δ on column 1. For reversible P, $\lambda \leq \gamma$ (Section A.1, T9); for directed P the leak can exceed the gap of S, as a machine-precision check shows.

Theorem 2.18 (Uniform distance law)**.** *Let P be row-stochastic, ν any stationary distribution (one always exists; no positivity is assumed), f a normalized amplifier with lottery K, and $\lambda = \lambda(P)$.*

(i) Law. $\|f(P) - \mathbf{1}\nu^\top\|_\infty \leq 2(1 - \nu_{\min})\,\mathbb{E}[(1-\lambda)^K]$.

(ii) Both ends attained. *On the two-community chain with $\nu = (\frac{1}{2}, \frac{1}{2})^\top$, the law is an equality for every lottery, with value $\mathbb{E}[(1-\gamma)^K]$. For strictly causal P with $\nu = e_1$ and $\lambda = \delta$, it is the envelope of Theorem 2.16(ii), attained by self-or-sink.*

(iii) The class extreme. *For $\lambda_0 \in (0, 1)$, the largest value of $\|f(P) - \mathbf{1}\nu^\top\|_\infty$ over operators with uniform leak at least λ_0 is exactly $2\mathbb{E}[(1-\lambda_0)^K]$, attained by self-or-sink at $\delta = \lambda_0$; a family of amplifiers keeps it bounded below if and only if its lotteries keep uniform short-path mass.*

The proof is in Section A.3. The engine is the classical one-step contraction $d_{\mathrm{TV}}(\mu P, \mu' P) \leq (1-\lambda) d_{\mathrm{TV}}(\mu, \mu')$ (Doeblin, 1937; Dobrushin, 1956; Seneta, 1981), run through the lottery. An exact identity for the distance, a multi-step version, two floors and a conditioning variant complete the law; their proofs are in Section A.3 as well.

Proposition 2.19 (Further clauses of the uniform law)**.** *In the setting of Theorem 2.18, with $\mathcal{L}_i$ the law of the lottery-stopped walk from token i and a the anchor:*

(i) Identity. $\|f(P) - \mathbf{1}\nu^\top\|_\infty = 2 \max_i d_{\mathrm{TV}}(\mathcal{L}_i, \nu)$*; for strictly causal P with $\nu = e_1$ this is Theorem 2.16(i).*

(ii) Multi-step law. *For every $m \geq 1$,* $\|f(P) - \mathbf{1}\nu^\top\|_\infty \leq 2(1 - \nu_{\min})\,\mathbb{E}\big[(1 - \lambda(P^m))^{\lfloor K/m \rfloor}\big]$.

(iii) Floors. $\|f(P) - \mathbf{1}\nu^\top\|_\infty \geq \max\{|f(\zeta)| : \zeta \in \mathrm{spec}(P),\ \zeta \neq 1\}$ *(with the maximum defined as zero for an empty set) and* $\geq 2(a - \nu_{\min})_+$.

(iv) Conditioning variant. *If $\pi > 0$, then for every $m \geq 1$,*

$$\|f(P) - \mathbf{1}\pi^\top\|_\infty \leq \sqrt{(1 - \pi_{\min})/\pi_{\min}}\ \mathbb{E}\big[\psi_m^{\lfloor K/m \rfloor}\big],$$

where ψ_m is the π-norm of P^m restricted to $\mathbf{1}^{\perp_\pi}$; for lazy reversible irreducible P and $m = 1$ this reads $\sqrt{(1 - \pi_{\min})/\pi_{\min}}\,\mathbb{E}[(1-\gamma)^K]$.

In a synthetic check, the law holds with no violation over 2,884 random operators from six families. Three remarks. *The prices of uniformity.* Relative to the two exact end laws, the uniform law pays a constant $2(1 - \nu_{\min}) \in [1, 2]$ and replaces the spectral gap by the leak; both prices are attained. *The leak is not a function of the reversible part.* On the lopsided triangle (Worked

example 1.7), P_s and its time reversal P_{-s} share S, the flow ratio and the spectrum, yet at $s = 0.1$ their leaks are 0.55 and 0.60 (a machine-precision check): the leak, which decides how fast content is destroyed in the ℓ_∞ reading, sees the direction of circulation, and on many operators so does the ℓ_∞ distance itself, which differs between $f(P)$ and $f(P^\dagger)$ in 142 of 200 random cases of a synthetic check. *An anchor floor that does not depend on ν lives at the ends.* Every operator obeys the floor $2(a - \nu_{\min})_+$ of Proposition 2.19, but it vanishes when $a \le \nu_{\min}$, and then an anchored lottery can cancel: for the two-token swap, a machine-precision check gives $f(P) = \frac{1}{2}(I + P) = \mathbf{1}\pi^\top$ exactly although $a = \frac{1}{2}$. At the lazy reversible end the distance is at least a in both norms, because every eigenvalue lies in $[0, 1]$ and so $f(\mu_2) \ge a$; at the causal end it is at least $2a$ (Theorem 2.16(iii)).

A trained network does not repeat one operator; it composes different ones, joined by a residual stream. Proposition 1.5 makes that composite a lottery over layer subsets, and the laws above carry over with one scalar per layer.

Proposition 2.20 (The residual-stack clock)**.** *Let $F = \prod_{\ell=L}^{1} ((1 - b_\ell)I + b_\ell P_\ell)$ with row-stochastic P_ℓ and update shares $b_\ell \in [0, 1]$.*

(i) Causal layers. *If every P_ℓ is strictly causal with sink weight δ_ℓ, then $2\prod_\ell(1 - b_\ell) \le \|F - \mathbf{1}e_1^\top\|_\infty \le 2\prod_\ell(1 - b_\ell\delta_\ell)$.*

(ii) A shared stationary distribution. *If every P_ℓ has stationary distribution ν, then $\|F - \mathbf{1}\nu^\top\|_\infty \le 2(1 - \nu_{\min})\prod_\ell (1 - b_\ell\lambda(P_\ell))$.*

(iii) A representation floor. *Let $X_{\ell+1} = (1 - b_\ell)X_\ell + b_\ell P_\ell X_\ell$, where each P_ℓ is row-stochastic and may be computed from X_ℓ, as attention is. With $\theta_\ell = \min_i (P_\ell)_{ii}$ and $\operatorname{diam}(X) = \max_{i,j}\|x_i - x_j\|$ in any norm, $\operatorname{diam}(X_{\ell+1}) \ge (1 - 2b_\ell(1 - \theta_\ell))\operatorname{diam}(X_\ell)$. For $b_\ell \le \frac{1}{2}$ the bound is attained by the two-token swap.*

The proof is in Section A.3. Clause (i) is the scalar-share case of the survival bounds of Section 8.3, which allow a different share for every token. The clock of a residual stack is therefore $\sum_\ell b_\ell\delta_\ell$ (or $\sum_\ell b_\ell\lambda_\ell$), not the leak of any single layer: a single captured causal attention layer, iterated, reaches half its distance to the first token within a median of one or two applications, while the distance of a composite of such layers is governed by how much of each layer the stream admits (Section 2.7, Figure 2.7b). Clauses (i) and (ii) concern the realized product and hold however the P_ℓ arose, including attention rebuilt from the state; clause (iii) bounds the representation itself along any trajectory of this simplified block. It is the quantitative form, for the block without value maps, of the observation of Dong et al. (2021) that skip connections prevent the collapse of attention-only networks, and it shows what the floor costs: for $0 \le b_\ell \le \frac{1}{2}$, the iterated floor $\prod_\ell(1 - 2b_\ell)$ decays with depth unless the shares are summable; lazy attention ($\theta_\ell \ge \frac{1}{2}$) improves the per-layer factor to $1 - b_\ell$ but does not stop the decay. The composite of lazy layers $\frac{1}{2}(I + A_\ell)$ is the attention rollout of Abnar & Zuidema (2020), the case $b_\ell = \frac{1}{2}$. Token-wise maps between layers (MLPs, normalization) preserve the consensus set but are outside the statement: a lower bound on the diameter would need a lower (co-Lipschitz) constant for them.

2.6 Normalization placement is a choice of budget

Normalization decides how a branch's weight relative to the stream evolves with depth, and hence the budget. In Post-LN, $X \leftarrow \mathrm{Norm}(X + \mathrm{branch})$ resets the stream after each layer. Every layer then keeps the same relative weight, the budget grows linearly, and forgetting is exponential. In Pre-LN, $X \leftarrow X + \mathrm{branch}(\mathrm{Norm}(X))$, each branch adds an increment of fixed size while the stream accumulates all previous increments. For incoherent (centered) increments, the relative energy of

layer ℓ then decays like $1/\ell$ and its relative amplitude like $\ell^{-1/2}$ (Liu et al., 2020a; Xiong et al., 2020; De & Smith, 2020). The energy law makes the consequence exact.

Proposition 2.21 (Pre-LN sits on the critical line)**.** *In the energy law without drift, let each token-mixing branch add expected energy v to the consensus channel and each token-wise branch ($P = I$, such as an MLP) add v', independently of the stream, and let the branch strength be set by the consensus energy $E_{\ell-1}$ (the mean-field form of Pre-LN). Then $\alpha_\ell^2\sigma^2 = v/E_{\ell-1}$ with $E_\ell = E_0 + \ell(v+v')$, and with the attention share $s = v/(v+v')$ and $e = E_0/(v+v')$,*

$$R_L(\mu) = \prod_{\ell=1}^{L} \frac{e+\ell-1+s\mu^2}{e+\ell-1+s} = \frac{\Gamma(e+L+s\mu^2)\,\Gamma(e+s)}{\Gamma(e+s\mu^2)\,\Gamma(e+L+s)} \;\sim\; C\,L^{-s(1-\mu^2)}. \tag{2.9}$$

The expected number of layers that mix in $(aL, bL]$ converges to $s\ln(b/a)$.

The proof, with the other placements, is in Section A.3. Pre-LN is critical: its budget $s\log L$ sits on the boundary between a floor and exponential loss, in the incoherent regime and, with amplitudes in place of energies, in the coherent regime, where the amplitude ratio is $\asymp L^{-s(1-\mu)}$. The share profile alone does not decide the regime; coherence does. Under Proposition 2.21 the branch weight falls like $\ell^{-1/2}$ and the loss is polynomial, whereas the same profile with a fixed positive drift forgets as $\exp(-\Theta(\sqrt{L}))$ (the $c/\sqrt{\ell}$ row of Table 2.2). A self-normalized branch with a drift, however, grows the stream linearly, so its amplitude share falls like $1/\ell$ and relative energy decays as $L^{-2s(1-\mu)}$, again polynomially (Section A.3). In the large-depth limit, layers $L/2$ to L contribute $s\ln 2 \approx 0.69\,s$ expected mixing layers: mixing is spread log-uniformly over depth. This is an initialization-level account of the observation that the deeper half of Pre-LN language models changes their computation little (Csordás et al., 2025; Gromov et al., 2025; Men et al., 2025). The same law places the remedies proposed for this effect. For incoherent updates, scaling each branch by $c/\sqrt{L}$, as in GPT-2's initialization and depth-μP (Radford et al., 2019; Yang et al., 2024a), bounds the budget and restores a floor. LayerNorm Scaling multiplies the normalized input of layer ℓ by $1/\sqrt{\ell}$ (Sun et al., 2025); the stream's expected consensus energy then grows only like $\log\ell$, the budget like $\log\log L$, and relative energy decays as $(\log L)^{-s(1-\mu^2)}$, far more slowly than under Pre-LN but still without a floor. Table 2.2 collects these cases. The individual ingredients are known: the $1/\ell$ energy share, linear growth of the stream's energy, and polynomial rather than exponential clustering (Liu et al., 2020a; De & Smith, 2020; Karagodin et al., 2025). What is new is one criterion covering all rows, the exact exponent $s(1-\mu^2)$, and the log-uniform allocation. It also resolves an apparent conflict: Karagodin et al. (2025) find that a $1/\sqrt{t}$ layer-index scaling clusters faster than polynomially, whereas Theorem 2.13 gives polynomial loss for $1/\sqrt{\ell}$. Their dynamics are coherent, with budget $\sum\alpha_\ell \asymp \sqrt{L}$; those of Theorem 2.13 are incoherent, with budget $\sum\alpha_\ell^2 \asymp \log L$. Linear Pre-LN stacks reproduce the exponents across the whole spectrum (Figure 2.2c), and the measurements at the end of this section test Proposition 2.21 on randomly initialized transformers.

Read as mixing budgets, the residual recipes in use fall into the rows of Table 2.2. Each recipe is a branch weight α_ℓ relative to the stream; its budget is $\asymp \sum_\ell \alpha_\ell$ when updates add coherently and $\asymp \sum_\ell \alpha_\ell^2$ when they are incoherent, and the budget fixes how fast relative modal energy is lost.

Randomly initialized transformers test Proposition 2.21. Transformer blocks with independent Gaussian weights are stacked on 64 tokens; each block has softmax attention (4 heads at width 128 and 8 at width 512), a GELU MLP and LayerNorm without learned parameters, and each layer draws fresh weights. The inputs have i.i.d. standard Gaussian entries plus a shared Gaussian offset of scale 2, so the consensus component starts larger than the token distinctions. The placements are Pre-LN, Post-LN, Pre-LN with both branches scaled by $1/\sqrt{2L}$ or $1/\sqrt{L}$ for a network of total depth L, and Pre-LN with the normalized input of layer ℓ scaled by $1/\sqrt{\ell}$. The statistic is the

Table 2.2: **Residual recipes are mixing budgets.** Branch weight α_ℓ relative to the stream, and the resulting loss of relative modal energy for coherent updates (budget $\asymp \sum_\ell \alpha_\ell$) and incoherent updates (budget $\asymp \sum_\ell \alpha_\ell^2$). "Floor" means a positive limit; s is the attention share of branch energy (of branch amplitude in the coherent column).

Branch weight α_ℓ	Coherent	Incoherent	Examples
constant	$e^{-\Theta(L)}$	$e^{-\Theta(L)}$	Post-LN; unscaled residuals
$c/\sqrt{L}$ (total depth)	$e^{-\Theta(\sqrt{L})}$	floor	GPT-2 init., depth-μP, DeepScaleLM, DeepNet
c/L (total depth)	floor	$R_L \to 1$	CompleteP, nGPT
$c/\sqrt{\ell}$ (layer index)	$e^{-\Theta(\sqrt{L})}$	$L^{-c^2\sigma^2(1-\mu^2)}$	DS-Init; sqrt-scaling
c/ℓ (layer index)	$L^{-2c(1-\mu)}$	floor	—
Pre-LN (implicit)	$L^{-2s(1-\mu)}$	$L^{-s(1-\mu^2)}$	GPT-2, Pythia, Llama-style blocks
Pre-LN with LayerNorm Scaling	—	$(\log L)^{-s(1-\mu^2)}$	Sun et al. (2025)

energy of token distinctions relative to the mean token, $R(X) = \|X - \mathbf{1}m^\top\|_F^2/(N\|m\|^2)$ with m the mean token, divided by its value at the input; curves are geometric means over three seeds. At initialization the head-averaged bidirectional attention is nearly uniform. For its symmetrized form, the mean squared eigenvalue after excluding the leading eigenvalue is below 1.3×10^{-5}, averaged over the last seven eighths of the layers and three seeds. Proposition 2.21 then predicts a power law with exponent $-s$, where s is the attention share of the branch energy written into the consensus channel. Figure 2.3a shows the four regimes of Table 2.2 in one network. Post-LN keeps 4.5×10^{-3} of the distinction energy after 8 layers and about 5.7×10^{-10} after 32. Pre-LN decays polynomially, with fitted slope -0.69 against the predicted $-s = -0.70$. Scaling both branches by $1/\sqrt{2L}$, GPT-2's initialization, keeps 0.91–0.96 at the four tested total depths from 8 to 512, and $1/\sqrt{L}$ keeps 0.83–0.90 at those depths (Table 2.4). LayerNorm Scaling keeps 0.50 at $L = 512$ and is still declining, as the $(\log L)^{-s}$ law predicts. Varying the MLP's strength moves s from 0.2 to 1, and the fitted exponent follows $-s$ (Figure 2.3b, Table 2.3): at width 512 the largest deviation over seven configurations is 0.041. Attention-only stacks give -0.99 and -0.98 against the predicted -1. The one visible discrepancy, a GELU MLP at double strength and width 128, shrinks from 0.10 to 0.03 at width 512; there the nonlinear MLP slightly amplifies token distinctions relative to consensus, and finite width magnifies the effect. Linear token-mixing stacks with per-token RMS normalization test the μ-dependence (Figure 2.2c, Table 2.5): at width 512, coherent exponents match $-2(1-\mu)$ to within 0.001 across the eleven sampled modes, and incoherent ones match $-(1-\mu^2)$ to within 0.05. The incoherent discrepancies come from per-token normalization, which the mean-field law replaces by its expectation; they shrink from width 128 to 512. Under causal masks the decay stays polynomial with smaller exponents: fitted slopes lie between -0.14 and -0.68, with magnitudes short of s (ratios 0.5–0.9). The difference is attributed to the causal kernels, to which Proposition 2.21 does not apply directly: causal attention collapses toward the first token rather than the mean, near-uniform causal attention has non-consensus eigenvalues $1/2, 1/3, \ldots$ rather than values near zero, and its kernels are far from normal, so the energy of a sum of modes is not the sum of modal energies.

2.7 Measured: graphs and pretrained language models

On graphs, the measurements use the largest connected component of Cora (Yang et al., 2016) (2485 nodes, 5069 edges) and the duplicate-filtered Chameleon graph (Platonov et al., 2023) (890 nodes, 8854 edges), with the symmetric self-loop operator $H = \widetilde{D}^{-1/2}(W+I)\widetilde{D}^{-1/2}$, which is similar

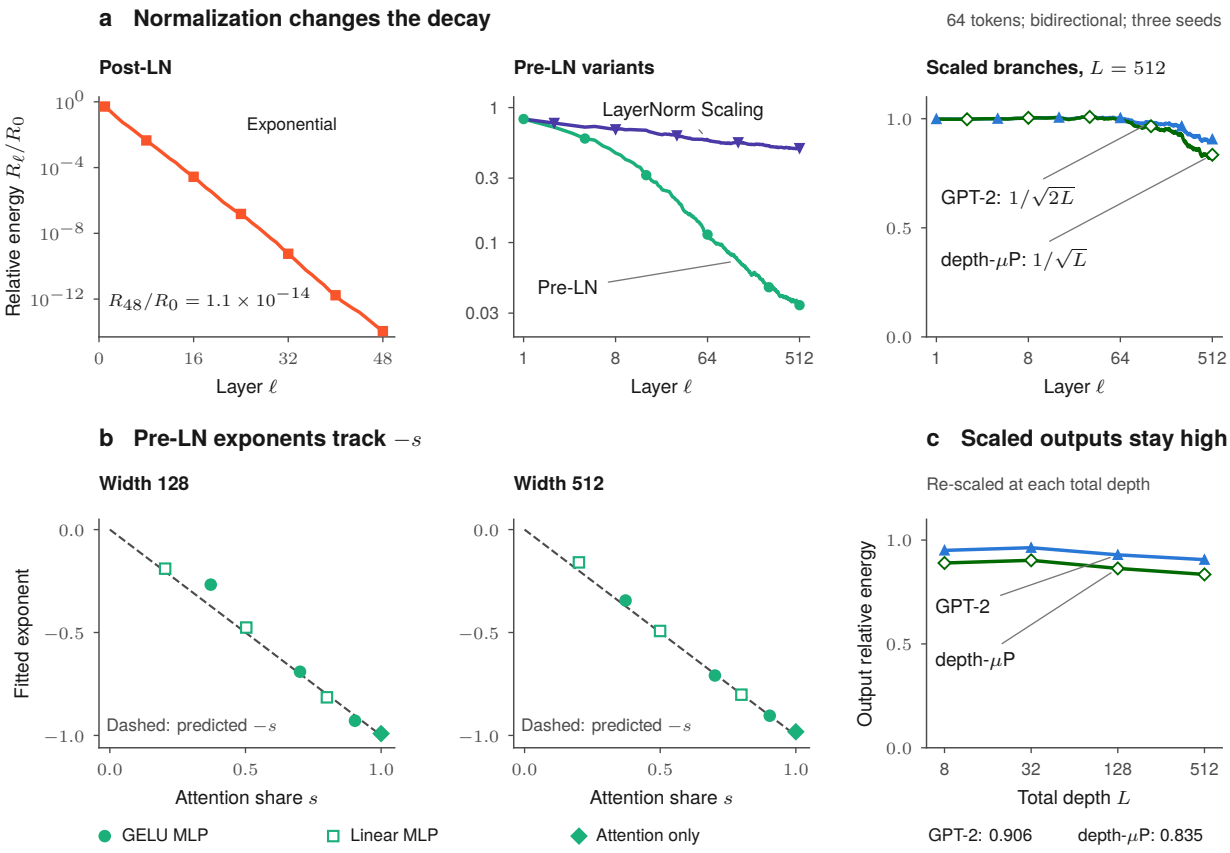


Figure 2.3: **Normalization placement changes the decay rate at initialization, and depth-scaled branches retain high output energy.** (a) Every saved positive layer of five width-128 placements, plotting token-distinction energy relative to consensus and normalized at the input, R_ℓ/R_0. The facets use different scales: Post-LN has a logarithmic ordinate and linear layer axis; Pre-LN and LayerNorm Scaling use logarithmic axes; scaled branches have a linear ordinate and logarithmic layer axis. The two scaled trajectories hold total depth fixed at $L = 512$, with both branches multiplied by the labelled factors. Symbols mark actual layers without offsets. (b) All fourteen bidirectional Pre-LN fits, separated by width on identical axes. Circles use GELU, open squares a linear MLP, and diamonds attention only. The exponent is fitted to log energy at layers $32, 64, 128, 256, 512$; s is the measured attention share of branch energy written into consensus. The dashed $-s$ is the mean-field prediction of Proposition 2.21. (c) Distinct depth-family measurements: output energy from a newly instantiated network at each total L, with its branches scaled for that L; the numbers below are the $L = 512$ outputs. All networks use 64 tokens and three seeds per configuration. Energies are geometric means and s is an arithmetic mean; per-seed spread was not saved, so no intervals can be shown. Records: `results/transformer_init.json` (Bruce, 2026k).

Table 2.3: Fitted Pre-LN exponents at initialization. The MLP multiplier κ varies the attention share s, measured as the attention branch's fraction of the energy that both branches write into the mean token, averaged over the last seven eighths of the layers; the linear MLP replaces GELU by the identity, so its token-wise branch is exactly the $P = I$ branch of Proposition 2.21. Exponents are least-squares slopes of $\log R$ against $\log L$ on $L \in \{32, 64, 128, 256, 512\}$. "Deviation" is $|\text{slope} + s|$ for bidirectional attention.

mask	width	MLP	κ	share s	fitted slope	$-s$	deviation
bidirectional	128	none	0	1.000	-0.991	-1.000	0.009
bidirectional	128	GELU	0.5	0.904	-0.929	-0.904	0.025
bidirectional	128	GELU	1	0.701	-0.691	-0.701	0.010
bidirectional	128	GELU	2	0.372	-0.267	-0.372	0.104
bidirectional	128	linear	0.5	0.800	-0.815	-0.800	0.015
bidirectional	128	linear	1	0.503	-0.476	-0.503	0.027
bidirectional	128	linear	2	0.204	-0.190	-0.204	0.014
bidirectional	512	none	0	1.000	-0.982	-1.000	0.018
bidirectional	512	GELU	0.5	0.904	-0.904	-0.904	0.000
bidirectional	512	GELU	1	0.702	-0.709	-0.702	0.007
bidirectional	512	GELU	2	0.372	-0.345	-0.372	0.028
bidirectional	512	linear	0.5	0.799	-0.802	-0.799	0.003
bidirectional	512	linear	1	0.500	-0.493	-0.500	0.007
bidirectional	512	linear	2	0.200	-0.159	-0.200	0.041
causal	128	none	0	1.000	-0.636		
causal	128	GELU	0.5	0.903	-0.677		
causal	128	GELU	1	0.700	-0.494		
causal	128	GELU	2	0.368	-0.190		
causal	128	linear	0.5	0.800	-0.617		
causal	128	linear	1	0.502	-0.379		
causal	128	linear	2	0.202	-0.186		
causal	512	none	0	1.000	-0.624		
causal	512	GELU	0.5	0.903	-0.615		
causal	512	GELU	1	0.700	-0.535		
causal	512	GELU	2	0.370	-0.290		
causal	512	linear	0.5	0.798	-0.580		
causal	512	linear	1	0.497	-0.419		
causal	512	linear	2	0.199	-0.142		

Table 2.4: Relative energy of token distinctions at the output of depth-scaled Pre-LN networks (width 128, three seeds).

total depth L	8	32	128	512
branch $\times 1/\sqrt{2L}$ (GPT-2)	0.950	0.963	0.929	0.906
branch $\times 1/\sqrt{L}$ (depth-μP)	0.890	0.903	0.864	0.835

Table 2.5: Linear Pre-LN stacks (width 512): predicted and fitted exponents of relative modal energy. The operator is the lazy ring walk $P = \frac{1}{2}I + \frac{1}{4}(Z + Z^\top)$ on 32 tokens, with Z the cyclic shift, whose eigenvalues $\frac{1}{2} + \frac{1}{2}\cos(2\pi k/32)$ span $[0, 1]$. Incoherent branches use fresh V_ℓ with entries $\mathcal{N}(0, 1/d)$; coherent branches use $V_\ell = I$; the variant with a token-wise branch alternates a token-mixing and a token-wise branch, so $s = \frac{1}{2}$. Energies of eleven modes relative to consensus are averaged in log space over eight seeds (four for the coherent case), and slopes are fitted on $L \in \{128, \ldots, 1024\}$. The Post-LN columns compare the fitted per-layer factor with $(1 + \mu^2)/2$; LNS is Pre-LN with LayerNorm Scaling.

	Pre-LN, incoherent		Pre-LN, coherent		with token-wise branch		Post-LN factor		LNS
μ	pred.	fit	pred.	fit	pred.	fit	pred.	fit	fit
-0.000	-1.000	-0.981	-2.000	-2.000	-0.500	-0.495	0.500	0.501	-0.133
0.038	-0.999	-0.979	-1.924	-1.924	-0.499	-0.513	0.501	0.504	-0.133
0.084	-0.993	-0.972	-1.831	-1.831	-0.496	-0.482	0.504	0.515	-0.132
0.222	-0.951	-0.928	-1.556	-1.555	-0.475	-0.510	0.525	0.575	-0.126
0.309	-0.905	-0.878	-1.383	-1.382	-0.452	-0.455	0.548	0.613	-0.118
0.500	-0.750	-0.721	-1.000	-0.999	-0.375	-0.352	0.625	0.682	-0.097
0.598	-0.643	-0.626	-0.805	-0.804	-0.321	-0.328	0.679	0.718	-0.083
0.778	-0.395	-0.360	-0.444	-0.444	-0.198	-0.195	0.802	0.809	-0.052
0.854	-0.271	-0.255	-0.293	-0.292	-0.136	-0.126	0.864	0.864	-0.031
0.962	-0.075	-0.033	-0.076	-0.076	-0.037	-0.021	0.963	0.965	-0.018
0.990	-0.019	0.015	-0.019	-0.019	-0.010	0.002	0.990	0.990	-0.006

to a stochastic walk. H is diagonalized completely. Centering the one-hot labels Y removes the consensus direction: $Y_c = (I - uu^\top)Y$, where $u = \widetilde{D}^{1/2}\mathbf{1}/\|\widetilde{D}^{1/2}\mathbf{1}\|$. The retained label energy of a filter is the share of Y_c that it keeps, $R_Y(f) = \|f(H)Y_c\|_F^2/\|Y_c\|_F^2 = \sum_j e_j f(\mu_j)^2 / \sum_j e_j$, where e_j is the label energy in eigenvector j. The two graphs place the signal very differently (Figure 2.4a, Table 2.7): 66% of Cora's label energy lies in eigenvalues above 0.9 and 8.6% in negative ones, while filtered Chameleon has 11% above 0.9 and 34% below zero. Their normalized label Dirichlet energies, the label-weighted averages of $1 - \mu$, are 0.24 and 0.78. The head decides retention. At depth 64, powers, a lazy residual and restart keep 0.072, 0.169 and 0.455 of Cora's label energy, and every law without a bounded window keeps losing it with depth while restart holds its value (Figure 2.4b,c, Table 2.8). Among the three equal-mean laws of Figure 2.1b,c, the persistent short branch keeps 0.185 at $L = 256$ against 0.005 for the other two. Matching one modal gain does not match the whole signal: calibrating five laws to $f(0.85) = \frac{1}{2}$ without looking at either graph, retained energy ranges from 0.41 (fractional diffusion) to 0.53 (lazy residual and heat) on Cora, and the geometric law keeps less than heat on Cora (0.506 against 0.533) but more on Chameleon (0.123 against 0.108) (Table 2.9; Table 2.10 compares diffusion laws at explicit budgets). A retention target must be paired with the signal's spectrum.

Accuracy follows the functional that the readout uses. We run label propagation from 20 training labels per class (Cora; the ten standard splits for Chameleon) and logistic-regression probes on propagated features $f(H)X$, for each law at depths 1 to 256 (Figure 2.5, Tables 2.12 and 2.13). Four predictions of Section 2.2 hold on Cora. Laws that escape collapse: at $L = 256$ label-propagation accuracy is 0.31 for powers and 0.57 for the lazy residual, against 0.74 for restart. The mean does not matter: the persistent and shrinking branches, both of mean L, reach linear-probe accuracies of 0.75 and 0.54 at $L = 256$. Retention without reach plateaus at one hop: the persistent branch keeps $f(H) \approx \frac{1}{2}H$ plus consensus, and its label propagation falls to the one-hop accuracy of 0.29–0.31 even though its feature probe stays high. Shape outlives retention: uniform averaging keeps accuracy

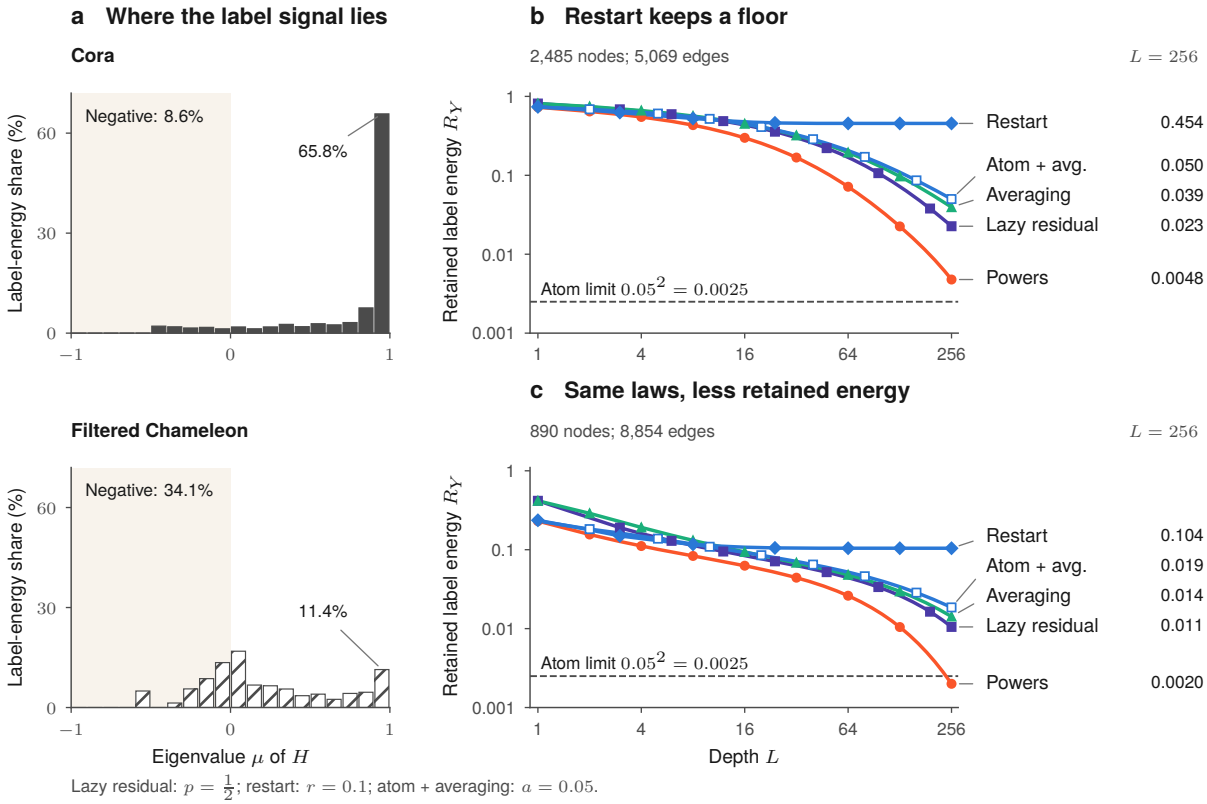


Figure 2.4: **Restart and a fixed input atom retain label energy; the graph spectrum determines how much.** (a) All twenty equal-width bins of centered label energy on the complete spectra, with common axes. Shading marks $\mu < 0$; annotations give the negative-mode energy share ($\mu < -10^{-12}$) and the share in the last bin ($\mu \geq 0.9$). (b, c) Retained energy $R_Y = \sum_{j\geq 2} e_j f(\mu_j)^2 / \sum_{j\geq 2} e_j$ for the same five laws on Cora and filtered Chameleon, on common logarithmic axes. Lines join every saved integer depth 1–256; symbols mark selected actual depths, with no coordinate offsets. The lazy residual and averaging coincide at $L = 1$. The right-hand values are the $L = 256$ samples. Parameters are $p = \frac{1}{2}$, $r = 0.1$ and $a = 0.05$; the atom law is $a + (1-a)L^{-1}\sum_{k=1}^{L} \mu^k$. Its dashed reference is the analytic energy limit $a^2 = 0.0025$, not a universal finite-depth bound on negative modes. Restart's spectrum-weighted limits are 0.454461 and 0.104397, respectively. These are deterministic calculations on fixed 2,485- and 890-node operators, without sampling uncertainty; every law has $R_Y(0) = 1$. Records: `results/graphs/summary.json`, `retention_curves.json` and both saved spectra (Bruce, 2026k).

Table 2.6: **Error decomposition by design class** (normalized denoising error; lower is better). "Best named" tunes powers, PPNP, heat and uniform averaging over their parameter; the nonnegative optimum is exact over all path laws on $\{0, \dots, 128\}$; "signed" is the best polynomial of degree 128 with $f(1) = 1$; the Wiener floor assigns each eigenvector its own gain.

Graph	noise	best named	nonneg. optimum	even depths only	signed	Wiener	optimal support
Cora	0.25×	0.148	0.142	0.142	0.137	0.124	$\{0, 2, 6\}$
	1×	0.281	0.274	0.274	0.262	0.242	$\{0, 2, 6\}$
	4×	0.420	0.413	0.413	0.395	0.376	$\{0, 2, 9, 10\}$
Chameleon	0.25×	0.201	0.194	0.194	0.190	0.170	$\{0, 2, 94, 95\}$
	1×	0.497	0.471	0.471	0.450	0.401	$\{0, 2, 58, 59\}$
	4×	0.789	0.752	0.752	0.711	0.666	$\{0, 2, 55, 56\}$

Table 2.7: Spectral summary of the two graphs. Negative modes are eigenvalues below -10^{-12}.

	Cora	Chameleon (filtered)
nodes / undirected edges	2485 / 5069	890 / 8854
raw nodes / components	2708 / 78	890 / 1
classes	7	5
μ_2 / relaxation time $1/(1-\mu_2)$	0.9964 / 276	0.9942 / 172
smallest eigenvalue	-0.48263	-0.59119
negative modes / their label-energy share	707 / 0.086	291 / 0.341
label-energy share with $\mu \geq 0.9$	0.658	0.114
normalized label Dirichlet energy	0.240	0.782
modes holding half the label energy	25	395

0.71 at $L = 256$ while its consensus-relative retention falls to 0.04, while Corollary 2.4 keeps its non-consensus modal gains in bounded ratios. Retained energy by itself is therefore a weak predictor of accuracy: a common modal target equalizes label-propagation accuracy to within 0.01 while retained energy varies by a third. The path law says instead which functional to read for which readout. On Chameleon, whose signal lies across the spectrum, every tested nonnegative law stays within 0.27–0.42. To see how much the class of path laws can do, we solve for the optimal law exactly. With the identity readout of Section 2.8 and white noise at 0.25, 1 and 4 times the per-node signal energy, the expected error of $f(H)$ is a convex quadratic in $(p_0, \dots, p_{128})$ on the simplex. Table 2.6 decomposes the error by design class. Tuned named laws (powers, PPNP, heat, averaging) come within 2–6% of the optimal nonnegative law; signed polynomials of degree 128 gain another 2–5%; and the per-mode Wiener floor lies 5–11% below the degree-128 signed optimum. The optimal laws have a telling shape: a zero-hop atom, mass at depth two, and at most one adjacent pair of deep powers, the two-point structure of Proposition 2.22. Depth one is always skipped, since it flips the sign of the negative-eigenvalue signal and, by Corollary 2.2(b), only even depths can pass that signal with the right sign. When we remove the negative-mode label energy from Cora, depth one returns to every optimum and odd depths carry 0.67–0.95 of the mass (Table 2.11).

The rollout path laws of sixteen pretrained causal language models come from three families: GPT-2 small to XL (12–48 layers) (Radford et al., 2019), Pythia from 70M to 6.9B parameters (6–32 layers), which share training data and data order (Biderman et al., 2023), and OPT from 125M to 6.7B (12–32 layers; OPT-350m is Post-LN) (Zhang et al., 2022). Each model is loaded at a pinned revision (Table 2.14) and run on the first 64 WikiText-103 validation paragraphs with at least 256 tokens (Merity et al., 2017). For every layer, input and query position the measurement records the

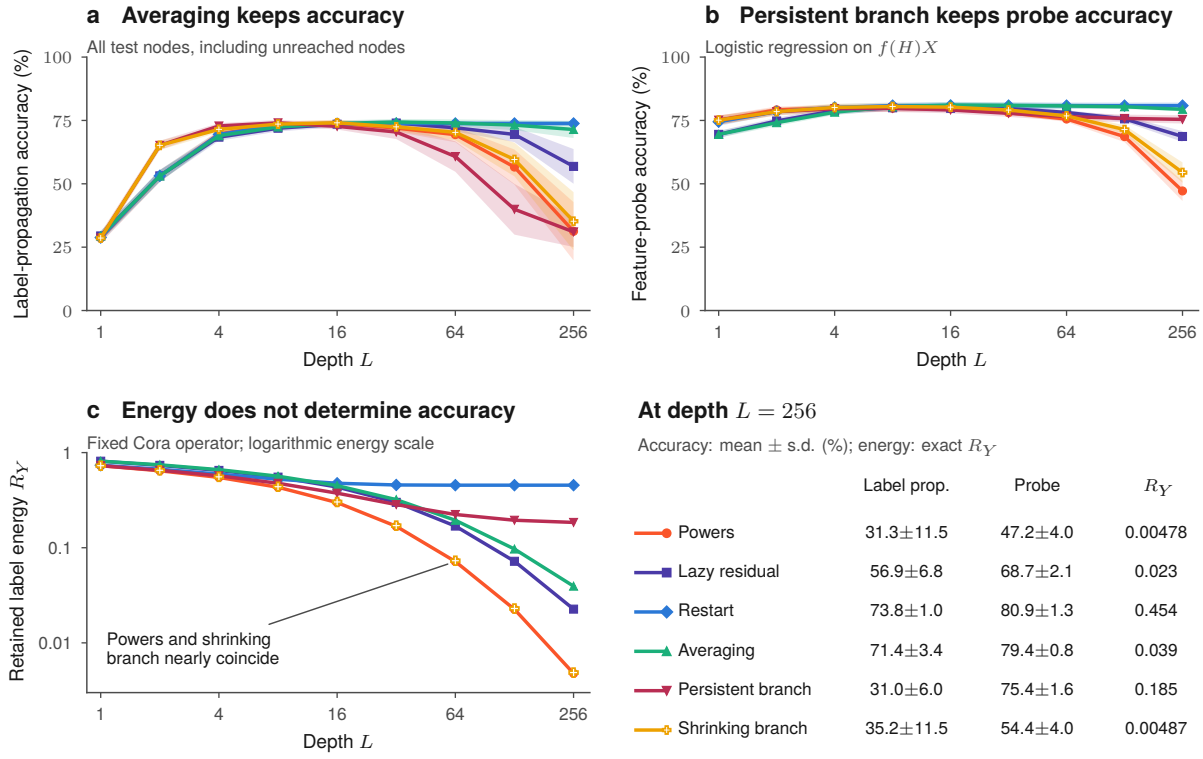


Figure 2.5: **Retention and task accuracy need not agree on Cora.** (a) Uniform averaging keeps label-propagation accuracy as depth grows. This is argmax accuracy over all test nodes, including nodes unreached by propagation, not a reached-node or abstention score. (b) The persistent short branch keeps feature-probe accuracy while its label propagation falls; the shrinking branch, with the same mean path length, loses probe accuracy. (c) Retained label energy R_Y for the same six laws: averaging loses energy despite its high accuracy. Open crosses show all nine recorded shrinking-branch values at their true coordinates over the nearly coincident powers curve; no positions are offset. The key lists all depth-256 outcomes. Accuracy bands and key entries show means ±1 population standard deviation over ten Cora splits (seeds 0–9, 20 training labels per class). Energy is an exact calculation on the fixed 2,485-node operator, without a sampling interval. Records: `results/graphs/task_level.json` (Bruce, 2026k); see Corollary 2.4 for shape retention.

Table 2.8: Retained centered label energy R_Y by depth.

	Cora				Chameleon			
L	4	16	64	256	4	16	64	256
powers	0.549	0.298	0.072	0.005	0.111	0.063	0.026	0.002
lazy residual, $p = \frac{1}{2}$	0.650	0.435	0.169	0.023	0.160	0.084	0.044	0.011
uniform averaging	0.662	0.450	0.194	0.039	0.192	0.094	0.048	0.014
atom 0.05 + averaging	0.629	0.447	0.206	0.050	0.148	0.092	0.052	0.019
restart, $r = 0.1$	0.595	0.476	0.455	0.454	0.138	0.108	0.104	0.104
shrinking short branch	0.556	0.303	0.073	0.005	0.119	0.064	0.026	0.002
persistent short branch	0.571	0.375	0.224	0.185	0.135	0.100	0.073	0.059
heat, $t = L/4$	0.750	0.556	0.302	0.072	0.306	0.116	0.063	0.026

Table 2.9: Five laws calibrated to the same modal gain $f(0.85) = \frac{1}{2}$ without reference to either graph: retained label energy and mean accuracy over ten splits (LP: label propagation; probe: logistic regression on $f(H)X$).

	Cora			Chameleon		
law calibrated to $f(0.85) = \frac{1}{2}$	R_Y	LP	probe	R_Y	LP	probe
residual (L=64, p=0.07181)	0.5333	0.724	0.802	0.1084	0.368	0.402
geometric / PPNP (r=0.13043)	0.5061	0.734	0.806	0.1228	0.363	0.360
heat (t=4.62098)	0.5328	0.724	0.802	0.1084	0.368	0.401
fractional (α=1/2, t=1.78970)	0.4115	0.733	0.794	0.1179	0.364	0.356
Matérn (ν=1.5, q=0.79658)	0.5152	0.732	0.807	0.1163	0.363	0.363

Table 2.10: Retained label energy for diffusion laws with explicit budgets. These fixed budgets are not claimed to be equivalent in computation or training cost.

law	Cora	Chameleon
DGC Euler (L=64, t=3)	0.5992	0.1348
heat (t=3)	0.5996	0.1354
fractional (t=3, α=1/2)	0.2733	0.0626
Matérn (ν=1.5, mean 3)	0.6205	0.1740

Table 2.11: Optimal nonnegative denoising laws by graph and noise level, against identity, the best tuned named law, even depths only, signed degree-128 polynomials and the per-mode Wiener floor. The last three rows are the control in which Cora's negative-mode label energy is set to zero before optimizing.

graph	noise	identity	best named	nonneg.	even only	signed	Wiener	optimal law (depth:weight)
Cora	0.25×	0.250	0.1484	0.1421	0.1421	0.1373	0.1236	0:0.456, 2:0.516, 6:0.028
Cora	1×	1.000	0.2815	0.2738	0.2738	0.2620	0.2418	0:0.177, 2:0.269, 6:0.554
Cora	4×	4.000	0.4195	0.4125	0.4126	0.3947	0.3757	0:0.058, 2:0.022, 9:0.657, 10:0.263
Chameleon	0.25×	0.250	0.2012	0.1941	0.1941	0.1898	0.1697	0:0.754, 2:0.222, 94:0.009, 95:0.015
Chameleon	1×	1.000	0.4972	0.4705	0.4705	0.4499	0.4010	0:0.430, 2:0.395, 58:0.071, 59:0.103
Chameleon	4×	4.000	0.7893	0.7520	0.7520	0.7110	0.6657	0:0.157, 2:0.291, 55:0.401, 56:0.151
Cora, negative energy removed	0.25×	0.250	0.1136	0.1113	0.1205	0.1058	0.0950	0:0.265, 1:0.666, 6:0.009, 7:0.059
Cora, negative energy removed	1×	1.000	0.2238	0.2216	0.2266	0.2146	0.1971	0:0.096, 1:0.288, 5:0.387, 6:0.229
Cora, negative energy removed	4×	4.000	0.3612	0.3551	0.3569	0.3420	0.3238	0:0.028, 1:0.075, 9:0.879, 10:0.018

Table 2.12: Cora: mean test accuracy over ten splits.

	label propagation					linear probe				
L	1	4	16	64	256	1	4	16	64	256
powers	0.286	0.696	0.739	0.694	0.313	0.751	0.798	0.797	0.755	0.472
lazy residual	0.294	0.684	0.735	0.721	0.569	0.695	0.788	0.798	0.782	0.687
averaging	0.294	0.690	0.738	0.740	0.714	0.695	0.783	0.811	0.807	0.794
atom + averaging	0.290	0.690	0.738	0.740	0.714	0.750	0.794	0.810	0.813	0.804
restart	0.288	0.695	0.737	0.738	0.738	0.745	0.801	0.809	0.809	0.809
shrinking branch	0.286	0.715	0.741	0.703	0.352	0.751	0.801	0.803	0.768	0.544
persistent branch	0.286	0.728	0.727	0.608	0.310	0.751	0.796	0.791	0.766	0.754
heat, $t = L/4$	0.698	0.704	0.721	0.740	0.696	0.631	0.734	0.802	0.799	0.757

Table 2.13: Filtered Chameleon: mean test accuracy over the ten standard splits.

	label propagation					linear probe				
L	1	4	16	64	256	1	4	16	64	256
powers	0.274	0.375	0.361	0.327	0.286	0.385	0.421	0.389	0.354	0.335
lazy residual	0.273	0.376	0.369	0.343	0.328	0.352	0.378	0.416	0.372	0.335
averaging	0.273	0.372	0.367	0.362	0.339	0.352	0.369	0.367	0.371	0.382
atom + averaging	0.273	0.372	0.367	0.362	0.339	0.368	0.379	0.366	0.347	0.334
restart	0.273	0.380	0.370	0.369	0.369	0.363	0.366	0.356	0.354	0.355
shrinking branch	0.274	0.367	0.361	0.329	0.290	0.385	0.398	0.404	0.373	0.336
persistent branch	0.274	0.343	0.333	0.303	0.279	0.385	0.385	0.385	0.376	0.372
heat, $t = L/4$	0.297	0.346	0.372	0.362	0.328	0.336	0.362	0.388	0.391	0.353

head-averaged attention row and the attention share $p_{\ell,i} = \|a_{\ell,i}\|/(\|a_{\ell,i}\| + \|x_{\ell-1,i}\|)$, where $a_{\ell,i}$ is the attention sublayer's output and $x_{\ell-1,i}$ the residual stream it is added to (Kobayashi et al., 2021). This defines a rollout (Section 8.3) with the network's own shares, and three budgets: all attention steps $\sum_\ell \bar{p}_\ell$, content-mixing steps $\sum_\ell \overline{p_\ell(1 - s_\ell - d_\ell)}$ that move mass to another content token, and sink absorptions $\sum_\ell \overline{p_\ell s_\ell}$, with s and d the attention paid to the sink and to the token itself and bars denoting averages over inputs and non-sink query positions. All survival and sink-treatment statistics are computed per input and then averaged. The self-weight means also exclude the first content position, which is a forced anchor.

- *Mixing grows much more slowly than depth.* Measured attention shares average 0.09–0.47 per layer, far below rollout's $\frac{1}{2}$, and they fall with depth: from shallowest to deepest model by a factor of 2.8 in GPT-2, 3.6 in Pythia and 1.7 in OPT (Table 2.15). The content-mixing budget grows much more slowly than depth. Fitting $S_L \propto L^\beta$ within each family gives $\beta = 0.07$ for Pythia, 0.38 for GPT-2 and 0.67 for OPT, against $\beta = 1$ for a fixed per-layer hazard (Table 2.16). Resampling the input sequences, paired across models because every model of a family saw the same texts, gives narrow 95% intervals: $[0.07, 0.08]$, $[0.37, 0.38]$ and $[0.66, 0.69]$. The larger uncertainty comes from having only four to seven models per family. A bootstrap over both models and inputs widens the intervals to $[-0.08, 0.16]$, $[0.27, 0.86]$ and $[0.62, 0.75]$, none of which reaches one; a t-interval over the four GPT-2 and five OPT models alone is wider still, $[-0.56, 1.31]$ and $[0.34, 1.00]$. Leaving out any one Pythia model gives slopes between 0.03 and 0.12. Width grows with depth in every family, and these models cannot separate the two: a width slope common to the families is -0.06 $[-0.31, 0.19]$, and adjusting for it gives depth slopes of 0.14, 0.41 and 0.76, with intervals that for OPT include one (Table 2.17). Pythia, the one suite trained on identical data in identical order, is the cleanest case, although it is also the only family here with parallel attention and MLP branches, so the two properties are confounded. A second parallel-block suite, Pythia trained on deduplicated data, reproduces the flat budget: its slope is 0.068 (interval over models $[-0.09, 0.23]$) against 0.075 for Pythia, and its budget is within 6.7% of Pythia's at every size (Bruce, 2026k). The ladder that would separate the two properties, sequential blocks trained on one corpus, could not be measured, because the Cerebras-GPT checkpoints planned for it are no longer publicly available; GPT-Neo's three models, sequential blocks trained on one corpus though not in one token order, give a slope of -0.27 with an interval over models of $[-1.25, 0.71]$, too few to decide. Pythia's seven models spend between 2.2 and 2.8 content-mixing steps per token from 6 to 32 layers; at fixed width 2048, going from 16 to 24 layers moves the budget from 2.40 to 2.42. The per-layer hazard profiles collapse when multiplied by L (Figure 2.6a), which is the signature of a depth family with bounded budget, the floor regime of Corollary 2.12. In the

language of Table 2.2, the trained Pythia models look less like Pre-LN at initialization, whose budget grows as $s \log L$, and more like the depth-scaled recipes, whose budget is fixed.

- *Which share, and which layers.* The norm share p is the leading-order clock when updates add coherently; when they are centered, energies add and the energy share $p_E = \|a\|^2/(\|a\|^2 + \|x\|^2)$ is the leading-order clock (Section 2.4). Neither is privileged, and for a trained model the exact consensus/contrast log of Section 2.9 decides between them. With p_E the budgets are smaller and dominated by a few layers. They do not grow with depth in Pythia or GPT-2 (slopes -0.27 and -0.20), whereas OPT's slope rises to 0.99, driven by its two smallest models; the intervals over models and inputs of GPT-2 and OPT then include one ($[-0.51, 1.08]$ and $[0.48, 1.17]$). Excluding the first layer, whose share reflects the small embedding stream, gives 0.07 $[-0.19, 0.21]$ for Pythia, 0.55 $[0.48, 0.83]$ for GPT-2 and 0.44 $[0.36, 0.64]$ for OPT, with intervals over models. Pythia's budget grows hardly at all under every definition; the GPT-2 and OPT slopes depend on the definition.

- *The sink absorbs depth.* What deeper Pythia models add is sink absorption: its budget rises from 0.25 steps per token at 6 layers to 1.7 at 32 (Figure 2.6b), and mean sink attention, excluding the first layer, rises from 0.14 to 0.60. Every attention step is a content move, a self-loop or a sink absorption. The extra attention steps that deeper Pythia models take are absorbed by the sink, while the content-mixing budget stays flat. Since, at fixed residual share and self-attention, attention to a zero-value sink cannot raise the mixing hazard (Section 8.3), this gives quantitative support to the view that sinks let deep models avoid over-mixing (Barbero et al., 2025a). The fitted sink budget rises with depth in all three families.

- *Rollout's convention overstates concentration.* Under the measured shares, the mean token keeps 14–76% of its rollout mass off the first token after all L layers; under $p = \frac{1}{2}$ it keeps less than 5% in fourteen of the sixteen, all but the two smallest Pythia models (Figure 2.6c). On earlier saved captures of batch-averaged attention products ($T = 64$, seven checkpoints), lowering p from $\frac{1}{2}$ to 0.1 delays the first depth at which worst-token survival reaches $\frac{1}{2}$ by a factor of two to three, or beyond the model's depth (Table 2.18). Claims that rollout concentrates on the first token "within a few layers" are claims about the convention. Two further checks close caveats of the earlier measurements. Under $p = \frac{1}{2}$, the product of batch-averaged kernels and the average of per-input products give the same mean survival to within 0.002 in fourteen of the sixteen models and within 0.04 in all of them. And the ordering of the sink treatments of Section 8.3 holds for every token of every input, with eviction leaving a mean self-weight 1.3 to 4.9 times smaller than a zero-value sink, a gap that generally widens with depth (Table 2.15). The family slopes are stable across context lengths: from 64 to 1024 tokens they stay within 0.37–0.38 for GPT-2, 0.08–0.16 for Pythia and 0.66–0.85 for OPT, provided that newline tokens, which become secondary sinks when paragraphs are joined (Xiao et al., 2024), are counted as sinks (Table 2.19).

At the causal end, two censuses measure the sink weight on trained models. Across 2,700 captured operators from seven autoregressive checkpoints at 64 tokens (2,544 heads and 156 head-averaged layers, each averaged over 16 inputs) (Bruce, 2026j) the median is 0.416 (0.410 per head), and 0.67% of the operators have $\delta < 10^{-6}$. On GPT-2 small (Radford et al., 2019), over sixteen WikiText-2 paragraphs (Merity et al., 2017) of 64 tokens, the 2,304 per-head operators have median 0.080 (interquartile range $[0.020, 0.216]$) with 2.2% below 10^{-6}, and the 192 head-averaged per-layer operators have median 0.29. The two censuses differ in models and protocol, and we do not attribute the difference; in both, a vanishing sink weight is rare, which is the regime in which the envelope of Theorem 2.16 bites. Iterating one captured layer tests the theorem on real instances of the operator class along an artificial perturbation, not on a trained trajectory: the median depth to half-collapse

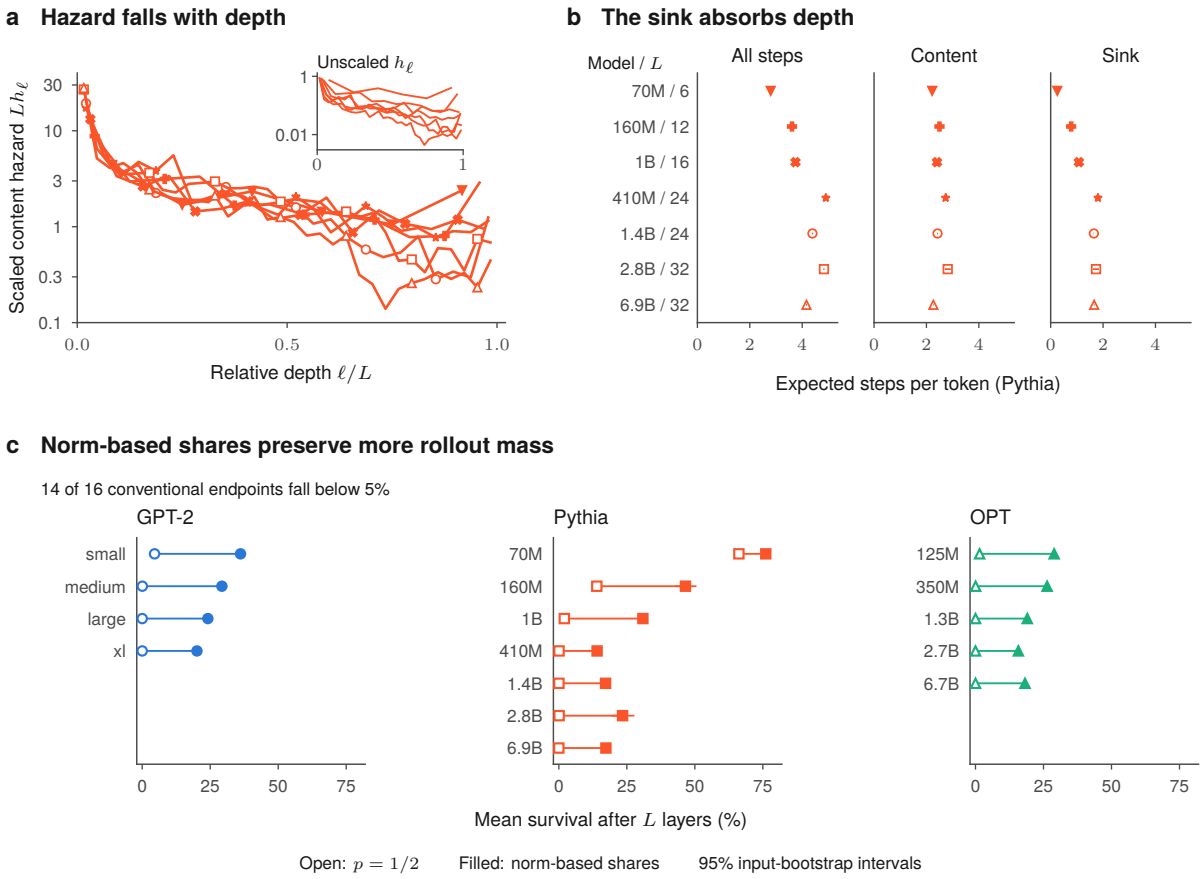


Figure 2.6: **Norm-based shares give Pythia a nearly flat content budget and preserve more rollout mass than the half-share convention.** (a) Content-mixing hazard h_ℓ multiplied by model depth L, with the unscaled hazard inset at the same relative depths. Both vertical axes are logarithmic and include the first-layer embedding-stream spike. (b) Seven Pythia models, labelled by parameter count and depth; marker shapes match (a). Separate axes with the same scale show all attention steps, content moves and sink absorptions; all steps equal content moves plus sink absorptions plus self-loops. Colour identifies model family throughout, not budget quantity. (c) All sixteen models' mean survival off the sink after their full L layers: filled marks use norm-based shares and open marks use $p = \frac{1}{2}$. Fourteen conventional endpoints are below 5%. Models are ordered by depth, then parameter count. Horizontal intervals in (b) and on filled marks in (c) are 95% input-bootstrap intervals, often smaller than the markers (64 WikiText-103 inputs at $T = 256$; `results/lm_summary_T256.json` (Bruce, 2026k)). These are rollout measurements, not interventions.

Table 2.14: Checkpoints and the Hugging Face revisions at which they were loaded. Pythia models use parallel attention and MLP branches; OPT-350m is the one Post-LN model.

model	revision	normalization	layers	width	heads	parameters
gpt2	`607a30d783`	pre-LN	12	768	12	124M
gpt2-medium	`6dcaa7a952`	pre-LN	24	1024	16	355M
gpt2-large	`32b71b1258`	pre-LN	36	1280	20	774M
gpt2-xl	`15ea56dee5`	pre-LN	48	1600	25	1,558M
pythia-70m	`a39f36b100`	pre-LN, parallel	6	512	8	70M
pythia-160m	`50f5173d93`	pre-LN, parallel	12	768	12	162M
pythia-1b	`f73d7dcc54`	pre-LN, parallel	16	2048	8	1,012M
pythia-410m	`9879c9b5f8`	pre-LN, parallel	24	1024	16	405M
pythia-1.4b	`fedc38a16e`	pre-LN, parallel	24	2048	16	1,415M
pythia-2.8b	`2a259cdd96`	pre-LN, parallel	32	2560	32	2,775M
pythia-6.9b	`c0e3eee36d`	pre-LN, parallel	32	4096	32	6,857M
opt-125m	`27dcfa74d3`	pre-LN	12	768	12	125M
opt-350m	`08ab08cc4b`	post-LN	24	1024	16	331M
opt-1.3b	`3f5c25d0bc`	pre-LN	24	2048	32	1,316M
opt-2.7b	`905a4b602c`	pre-LN	32	2560	32	2,652M
opt-6.7b	`a45aa65bbe`	pre-LN	32	4096	32	6,658M

Table 2.15: Measured path laws at $T = 256$ (64 inputs). Budgets are expected steps per token summed over layers; survival is the fraction of rollout mass off the first token after all L layers; self-weights are means over content positions, under the three sink treatments of Section 8.3.

		mean	budgets (steps per token)			survival after L layers		self-weight after L layers		
model	L	share	branch	content mixing [95% CI]	sink	measured: mean / worst	$p = \frac{1}{2}$: mean	skip F	zero value C	eviction X
gpt2	12	0.251	3.02	1.74 [1.72, 1.76]	0.93	0.362 / 0.474	0.0454	0.1306	0.1083	0.0450
gpt2-medium	24	0.120	2.88	1.59 [1.56, 1.61]	1.18	0.293 / 0.383	0.0005	0.2086	0.1931	0.0653
gpt2-large	36	0.124	4.46	2.80 [2.78, 2.83]	1.31	0.241 / 0.329	0.0000	0.0518	0.0471	0.0149
gpt2-xl	48	0.091	4.37	2.68 [2.65, 2.70]	1.47	0.201 / 0.363	0.0000	0.0722	0.0674	0.0178
pythia-70m	6	0.466	2.80	2.22 [2.20, 2.24]	0.25	0.761 / 0.947	0.6619	0.0490	0.0433	0.0337
pythia-160m	12	0.301	3.61	2.50 [2.44, 2.56]	0.78	0.466 / 0.686	0.1396	0.0454	0.0396	0.0184
pythia-1b	16	0.234	3.74	2.40 [2.38, 2.42]	1.08	0.309 / 0.419	0.0202	0.0380	0.0326	0.0125
pythia-410m	24	0.204	4.90	2.73 [2.71, 2.76]	1.80	0.141 / 0.209	0.0007	0.0393	0.0343	0.0077
pythia-1.4b	24	0.183	4.39	2.42 [2.40, 2.44]	1.65	0.172 / 0.244	0.0003	0.0416	0.0362	0.0095
pythia-2.8b	32	0.151	4.83	2.81 [2.70, 2.93]	1.73	0.234 / 0.420	0.0010	0.0314	0.0266	0.0058
pythia-6.9b	32	0.130	4.16	2.27 [2.25, 2.28]	1.66	0.173 / 0.256	0.0000	0.0355	0.0305	0.0080
opt-125m	12	0.194	2.33	1.08 [1.06, 1.09]	1.17	0.289 / 0.434	0.0152	0.3303	0.2985	0.0985
opt-350m	24	0.120	2.88	1.59 [1.58, 1.61]	1.18	0.263 / 0.366	0.0005	0.2185	0.2022	0.0662
opt-1.3b	24	0.150	3.59	1.93 [1.91, 1.94]	1.56	0.191 / 0.321	0.0002	0.0802	0.0678	0.0164
opt-2.7b	32	0.126	4.03	2.15 [2.12, 2.18]	1.78	0.158 / 0.305	0.0000	0.0583	0.0481	0.0097
opt-6.7b	32	0.117	3.74	2.01 [2.00, 2.03]	1.63	0.182 / 0.347	0.0000	0.0508	0.0398	0.0096

Table 2.16: Growth of budgets with depth within each family, $S_L \propto L^{\beta}$. A fixed per-layer hazard gives $\beta = 1$. The interval resamples input sequences independently for each model; Table 2.17 gives paired and model-level intervals.

family	layers	models	β, content mixing [95% CI]	β, branch	β, sink	content-mixing range
GPT-2	12–48	4	0.376 [0.366, 0.386]	0.316	0.328	1.59–2.80
Pythia	6–32	7	0.075 [0.061, 0.087]	0.298	1.138	2.22–2.81
OPT	12–32	5	0.673 [0.654, 0.690]	0.518	0.368	1.08–2.15

Table 2.17: Uncertainty of the family slopes of the content-mixing budget ($T = 256$, norm-based shares). *Inputs*: 95% bootstrap intervals over the input sequences, resampled independently for each model (unpaired, as in Table 2.16) or with one draw shared by all models of the family (paired). *Models*: the range of the slope when each model is left out in turn; a t-interval from the regression over models; and a bootstrap that resamples models and then inputs, redrawing samples with fewer than three distinct depths (their share is given as redrawn). The last column adds $\log d_{\text{model}}$ to a regression pooled over the sixteen models, with one width slope for all families, and gives t-intervals. The indented rows drop Pythia-6.9B, the model absent from the thirteen-model subset of Table 2.19, and OPT-350m, the one Post-LN model. The unpaired input bootstrap uses 4,000 resamples; the paired and model–input bootstraps use 10,000.

		Inputs		Models				
Family	$\hat{\beta}$	unpaired	paired	leave one out	t	with inputs	redrawn (%)	β given width
GPT-2	0.376	[0.366, 0.386]	[0.371, 0.380]	[0.309, 0.798]	[−0.555, 1.307]	[0.268, 0.858]	35	0.406 [0.096, 0.716]
Pythia	0.075	[0.061, 0.087]	[0.070, 0.079]	[0.033, 0.120]	[−0.070, 0.219]	[−0.077, 0.159]	3	0.135 [−0.187, 0.458]
without 6.9B	0.120	–	[0.111, 0.128]	[0.099, 0.140]	[0.009, 0.232]	[0.045, 0.229]	7	–
OPT	0.673	[0.654, 0.690]	[0.658, 0.687]	[0.597, 0.704]	[0.345, 1.001]	[0.616, 0.746]	47	0.758 [0.247, 1.268]
without 350m (Post-LN)	0.677	–	[0.662, 0.691]	[0.269, 0.729]	[0.258, 1.095]	[0.639, 0.750]	63	–

Width slope, common to the three families: −0.059 [−0.306, 0.188] (t-interval, 9 residual degrees of freedom).

Table 2.18: First depth at which worst-token rollout survival reaches $\frac{1}{2}$, by residual share, on the earlier batch-averaged captures. Here $> L$ means no crossing within the model's depth.

		residual share p							
checkpoint	L	0.05	0.1	0.15	0.2	0.3	0.5	0.75	1.0
gpt2	12	> 12	> 12	11	9	7	6	5	4
gpt2-medium	24	> 24	17	12	10	8	7	6	5
gpt2-large	36	33	21	17	15	12	10	8	6
pythia-160m	12	> 12	> 12	> 12	10	8	6	6	5
pythia-410m	24	> 24	17	14	12	10	8	7	7
pythia-1.4b	24	> 24	16	12	10	8	6	5	5
opt-350m	24	> 24	16	12	9	7	5	4	3

Table 2.19: Context-length robustness on the thirteen models run at every length (32 inputs per model; $T = 256$ uses 64). Family slopes β of the content-mixing budget, and the range of the content-mixing and sink budgets across models (steps per token). Inputs of 512 or 1024 tokens join consecutive paragraphs either with spaces or with blank lines; for the latter, both sink accountings are shown.

		slope β of content mixing				
input	T	GPT-2	Pythia	OPT	content mixing	sink
single paragraph	64	0.37	0.08	0.74	0.81–2.32	0.37–2.01
single paragraph	128	0.38	0.10	0.72	0.94–2.57	0.29–1.93
single paragraph (main)	256	0.38	0.12	0.73	1.08–2.81	0.25–1.80
paragraphs joined by spaces	512	0.38	0.14	0.75	1.15–3.23	0.17–1.78
paragraphs joined by spaces	1024	0.38	0.16	0.85	1.22–3.53	0.13–1.74
joined by blank lines; first-token sink	512	0.39	0.24	0.94	1.14–3.68	0.11–1.32
joined by blank lines; first token + newline sinks	512	0.37	0.13	0.67	1.12–2.90	0.31–1.90
joined by blank lines; first-token sink	1024	0.40	0.28	1.04	1.21–4.02	0.06–1.18
joined by blank lines; first token + newline sinks	1024	0.37	0.15	0.66	1.18–3.01	0.33–1.88

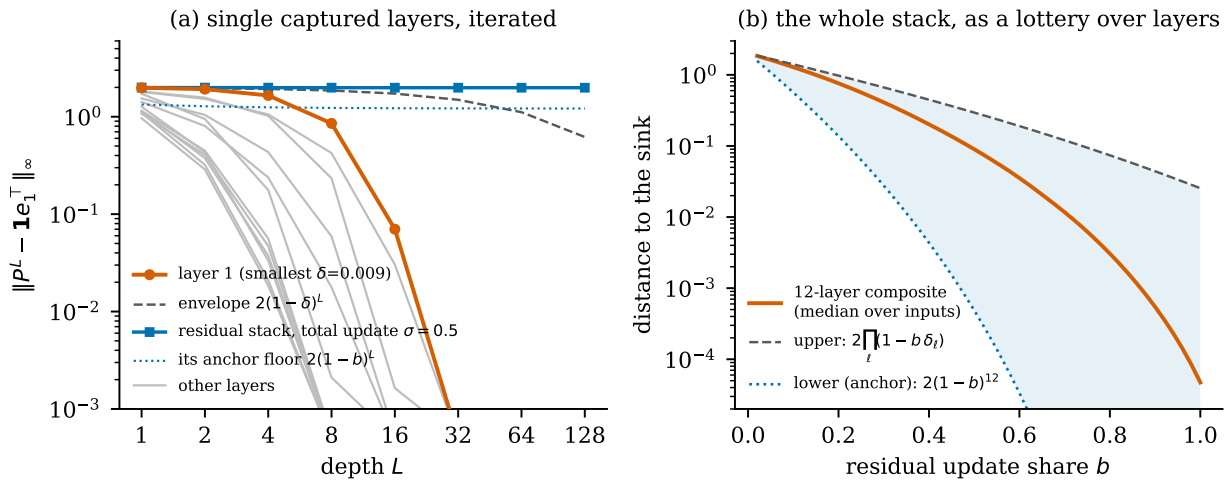


Figure 2.7: **The causal end on captured operators** (GPT-2 small, head-averaged per-layer attention on WikiText-2; real operators along an artificial perturbation). (a) For one input, each layer, iterated alone, collapses onto the first token inside the envelope $2(1-\delta)^L$ of Theorem 2.16; the highlighted layer has the smallest sink weight. A residual stack with total update $\sigma = 0.5$ ($b = \sigma/(L+\sigma)$ at depth L) holds its distance at every depth, above its anchor floor. (b) The twelve layers composed with a common update share b (Proposition 2.20(i)): the distance to the sink stays between the anchor bound $2(1-b)^{12}$ and the clock bound $2\prod_\ell(1-b\,\delta_\ell)$.

is one application (interquartile range $[1, 2]$) over 156 head-averaged layers of the seven checkpoints and two applications (maximum 8) on GPT-2. Composing GPT-2's twelve per-layer operators with a common update share b (Proposition 2.20(i)), again along an artificial perturbation of real operators, the median distance to the sink is 4.7×10^{-5} at $b = 1$ (attention alone), 0.090 at $b = \frac{1}{2}$, 0.76 at $b = 0.2$ and 1.28 at $b = 0.1$, out of a maximum of 2, always inside the two bounds of the proposition (Figure 2.7b). The share b is assumed here, not measured, and GPT-2's blocks carry value, output and MLP maps that the scalar-share model omits.

2.8 What no path law can do

Retention preserves distinctions; it does not decide which ones. A path law that keeps a useful mode also keeps every slower nuisance mode, and a readout sees both. A path law that retains more signal can therefore predict worse. In a four-state example (Section A.3) with signal and nuisance eigenvectors v_s and v_n, eigenvalues $\mu_s = \frac{19}{20}$ and $\mu_n = \frac{3}{10}$, input $x = v_s + a v_n$ with $a = \frac{3}{2}$, target v_s and identity readout, restart beats a lazy residual on signal and loses on error: at $L = 12$, the lazy residual $(\frac{1}{2} + \frac{z}{2})^{12}$ and the restart law with $r = \frac{1}{5}$ have signal gains 0.7380 and 0.8395, nuisance gains 0.0057 and 0.2632, and errors 0.0211 and 0.0559 in exact rational arithmetic; plain powers give an error of 0.0650. The example refutes a monotone relation between retention and fixed-readout error; it does not estimate how often the reversal occurs. Retention calculations are therefore paired with a statement of which modes carry signal. For nonnegative laws the trade-off between two modes has an exact limit.

Proposition 2.22 (The two-mode frontier of nonnegative filters)**.** *Fix $0 < \mu_- < \mu_+ < 1$ and a nuisance cap $\eta \in (0, 1)$. Choose k with $\mu_-^{k+1} \le \eta \le \mu_-^k$ and set $w = (\eta - \mu_-^{k+1})/(\mu_-^k - \mu_-^{k+1})$. Among*

all path laws with $f(\mu_-) \le \eta$, the largest possible signal gain is

$$\max f(\mu_+) = w\,\mu_+^k + (1-w)\,\mu_+^{k+1},$$

attained by the adjacent two-depth mixture $f(P) = wP^k + (1-w)P^{k+1}$.

Proof. The achievable points (μ_-^j, μ_+^j) lie on the concave curve $y = x^q$ with $q = \log \mu_+ / \log \mu_- \in (0,1)$. Their piecewise-linear interpolation U is concave and increasing, so $f(\mu_+) = \mathbb{E}[U(\mu_-^K)] \le U(\mathbb{E}\mu_-^K) \le U(\eta)$ by Jensen's inequality, and the two-point law attains both inequalities. □

Suppressing a mode at 0.5 to below 0.01 caps the gain at 0.9 at 0.4932, attained by $0.28P^6 + 0.72P^7$; no depth, restart probability or diffusion time does better, and a requested gain of 0.5 is infeasible. With more caps the problem is a linear program whose optimum mixes at most one more depth than there are active caps (Section A.3). On a lazy operator, whose spectrum lies in $[0,1]$, every path law is nondecreasing on the spectrum. The frontier then also caps every mode below μ_- and keeps every mode above μ_+, and it solves the band-pass problem exactly. With negative eigenvalues it need not; the optimal filter above has $f(-1) = -0.44$. Negative modes add their own caps, and Corollary 2.2(b) shows that even depths are the only nonnegative way to pass them with the right sign.

Leaving the class is how designs escape these limits. GPR-GNN and BernNet learn signed polynomial filters (Chien et al., 2021; He et al., 2021) and twicing attention uses $2A - A^2$ (Abdullaev & Nguyen, 2025): signed polynomials of one operator, which can exceed the frontier. The Differential Transformer subtracts two different attention maps (Ye et al., 2025) and leaves the class of combinations of one operator altogether, and learned feature maps with negative spectra can make graph dynamics dominated by high frequencies (Giovanni et al., 2023). Proposition 2.22 states what signed polynomials can buy, any gain above the frontier, and Section 2.7 measures how much that is worth on real label spectra; what the escape costs in coefficient variation is Proposition 2.9 (Section 2.3).

2.9 What holds in trained networks

Shown in an idealized setting. Depth as a distribution over path lengths (Sections 2.1 and 2.2): proved for path laws over powers of one fixed operator; exact for propagation on a fixed graph, not for the full nonlinear computation of a trained network (the stated scope of Bruce 2026k). The mixing budget as the rate of forgetting (Sections 2.4 and 2.6): stated for expected energies of residual stacks with random feature maps and mean-field normalization. Deep averaging forgets unless some paths stay short (Sections 2.2, 2.3 and 2.5): the if-and-only-if is proved for path laws over powers of one fixed operator. Bounds for realized products of distinct layers (Proposition 2.20) and a Dobrushin bound for sink-like attention (Section 5.5) carry its direction, that mixing without short paths forgets, to stacks whose operators change from layer to layer.

Measured in trained networks. In trained language models the sink budget rises with depth in all three families measured, and in Pythia the added depth goes to the sink while the content-mixing budget stays flat (Section 2.7): measured on rollout path laws, a linearization of attention flow built from each model's own attention and residual norms.

The exact path-law statements concern three objects: normalized nonnegative combinations of powers of one fixed operator (Sections 2.1 and 2.2 and the graph measurements of Section 2.7), expected energies of residual stacks with random feature maps, including the mean-field form of normalization (Sections 2.4 and 2.6 and the initialization measurements), and ordered products of stochastic kernels (Section 8.3). They do not describe the full nonlinear computation of a trained

network. Retention measures which distinctions survive; whether a task is solved is a separate question (Section 2.8). The path laws measured in pretrained models are rollout path laws built from each model's own attention and residual norms, an established linearization of attention flow, and the budgets are properties of that object. Within these boundaries the statements are exact; beyond them they are hypotheses that the framework makes precise and testable.

Section 21.1 collects the resulting design rules. The theory also suggests quantities worth logging. The per-layer attention share and sink mass, which cost one forward pass, give a network's mixing budget, and its growth with depth separates the regimes of Corollary 2.12. Retention itself can be logged exactly. Write the update of layer ℓ as $X_\ell = X_{\ell-1} + \Delta_\ell$, let $\Pi_0 = \mathbf{1}\mathbf{1}^\top/n$ and $\Pi_c = I - \Pi_0$ project onto the consensus and the contrast between tokens, and let $R_\ell = \|\Pi_c X_\ell\|^2/\|\Pi_0 X_\ell\|^2$. With $E_j = \|\Pi_j X_{\ell-1}\|^2$, $a_j = \langle \Pi_j X_{\ell-1}, \Pi_j \Delta_\ell\rangle/E_j$ and $b_j = \|\Pi_j \Delta_\ell\|^2/E_j$ for $j \in \{0, c\}$ at each layer,

$$\log \frac{R_L}{R_0} \;=\; -\sum_{\ell=1}^{L} \big[\log(1 + 2a_0 + b_0) - \log(1 + 2a_c + b_c)\big]$$

for any realized trajectory. The cross terms a_j are the coherent part of each update and the squares b_j the incoherent part, so the same log shows whether a model accumulates coherently, which by Theorem 2.13 decides the residual scaling it needs. Whether training actively regulates the budget, as the collapse of Pythia's hazard profiles suggests, and whether targeting a budget directly improves deep models, are questions the framework now makes measurable.

Uniform short-path mass is necessary and sufficient for a bounded collapse ratio at a positive content eigenvalue (Theorem 2.8), and for class-extremal distances bounded away from zero at the causal end and at every leak (Theorem 2.16(iii), Theorem 2.18(iii)). A positive anchor is its simplest sufficient form. At a fixed operator it gives the floor $2(a - \nu_{\min})_+$ (Proposition 2.19), which vanishes when $a \le \nu_{\min}$; a floor that does not depend on ν holds only at the ends, a at the lazy reversible end and $2a$ at the causal end, and in between an anchored lottery can cancel (Section 2.5). Filters with negative coefficients escape the characterization, and Proposition 1.4 says that escaping means ceasing to be an averaging filter. For placement, each order closes one channel of Theorem 4.1 and leaves the other open; neither is safe in general.

Chapter 3

Direction: preference, exchange and circulation

Section 1.2 split every walk into a reversible part S, which says who averages with whom, and a remainder J, which routes. Direction is the one property that the averaging alone cannot provide. A head in which every token attends to the same distribution q, so that $P = \mathbf{1}q^\top$, is as asymmetric as a stochastic matrix can be, yet it has no direction: no token precedes another, and the chain it defines is statistically identical when run backwards (Bruce, 2026g). This chapter locates direction exactly and shows why default statistics cannot see it. Chapter 0 states the finding in brief (Section 0.4.1), and Chapter 10 takes direction up as one of the conditions.

The split is the π-weighted one of Section 1.2, $P = S + J$ with $S = \frac{1}{2}(P + P^\dagger)$ and $J = \frac{1}{2}(P - P^\dagger)$ (1.2). The reversible part and the flow of Bruce (2026i) are the time-symmetric and time-antisymmetric parts of Bruce (2026g): the two splits are one decomposition, which the book writes with one set of symbols (Section R.4). Sections 3.1 to 3.4 develop it in each head's own stationary geometry. Sections 3.5 and 3.6 ask with what error the exact statements for reversible operators of Section 2.3 transfer to $P = S + J$, and the answer is a symmetry: reversing time flips the sign of the flow and leaves every quadratic readout unchanged. The circulation index of Section 3.2 is written κ_{circ}, to keep it apart from the collapse ratio κ_{col} of Chapter 2.

3.1 Attention as a stationary Markov operator

Let $P \in \mathbb{R}^{N\times N}$ be the post-softmax attention of one head on one input: $P_{ij} \geq 0$ is the weight with which reader token i attends to source token j, and $P\mathbf{1} = \mathbf{1}$. We assume a stationary law $\pi > 0$ with $\pi^\top P = \pi^\top$ and $\mathbf{1}^\top\pi = 1$. Softmax attention over a full token set has strictly positive entries, so π exists, is unique and is positive by the Perron–Frobenius theorem. Write $D_\pi = \mathrm{diag}(\pi)$ and $\langle f, g\rangle_\pi = f^\top D_\pi g$. The adjoint of P in $L^2(\pi)$,

$$P^\dagger = D_\pi^{-1}P^\top D_\pi, \qquad (P^\dagger)_{ij} = \frac{\pi_j P_{ji}}{\pi_i}, \tag{3.1}$$

is the *stationary time reversal*: if the chain generated by P is observed at stationarity and played backwards, it moves from i to j with probability $(P^\dagger)_{ij}$ (Aldous & Fill, 2002; Levin & Peres, 2017). It is itself row-stochastic with stationary law π. We set

$$\begin{gathered} S = \tfrac{1}{2}(P + P^\dagger), \qquad J = \tfrac{1}{2}(P - P^\dagger), \\ F = D_\pi P, \qquad A = \tfrac{1}{2}(F + F^\top) = D_\pi S, \qquad C = F - F^\top = 2D_\pi J. \end{gathered} \tag{3.2}$$

F_{ij} is the probability that a stationary step goes from reader i to source j. A is the *reciprocal traffic* on each pair of tokens, and C is the *current*: skew-symmetric with $C\mathbf{1} = 0$, a divergence-free flow on the token graph and hence a superposition of cycles (Schnakenberg, 1976; McCausland, 2007). The kernel is *reversible* (detailed balance) exactly when $J = 0$. Finally, stationary coordinates $\widetilde{M} = D_\pi^{1/2} M D_\pi^{-1/2}$ turn the $L^2(\pi)$ geometry into the Euclidean one. In them $\widetilde{S} = (\widetilde{P} + \widetilde{P}^\top)/2$ is symmetric and $\widetilde{J} = (\widetilde{P} - \widetilde{P}^\top)/2$ is skew. Edges are oriented from reader to source; the value update $(PV)_i = \sum_j P_{ij} V_j$ moves content the other way. Reversal, stilling (replacing P by S, Section 10.2) and all invariants below are defined for a frozen P captured from a forward pass. The stationary law is a reference that makes reversal a well-defined stochastic operation on that P. It is not an assumption that the network iterates the head to equilibrium.

3.2 Preference, exchange and circulation

We use the Hilbert–Schmidt norm $\|M\|_{\mathrm{HS}(\pi)} = \|\widetilde{M}\|_F$, the Frobenius norm in stationary coordinates. In these coordinates the split (3.2) becomes a sum of three orthogonal parts.

Theorem 3.1 (Preference, exchange, circulation)**.** *Let $u = \sqrt{\pi}$ (entrywise). Then u is a left and right eigenvector of $\widetilde{P}$ with eigenvalue one, and*

$$\widetilde{P} = \underbrace{uu^\top}_{\textit{preference}} + \underbrace{\widetilde{E}}_{\textit{exchange}} + \underbrace{\widetilde{J}}_{\textit{circulation}}, \qquad \widetilde{E} = \widetilde{S} - uu^\top,$$

where $\widetilde{E}$ is symmetric, $\widetilde{J}$ is skew and $\widetilde{E}u = \widetilde{J}u = 0$. The three terms are mutually orthogonal, so

$$\|\widetilde{P}\|_F^2 = 1 + \|\widetilde{E}\|_F^2 + \|\widetilde{J}\|_F^2, \qquad \|\widetilde{S}\|_F^2 - \|\widetilde{J}\|_F^2 = \mathrm{tr}(P^2). \tag{3.3}$$

Consequently the circulation index

$$\kappa_{\mathrm{circ}}(P) = \frac{2\|\widetilde{J}\|_F^2}{\|\widetilde{P}\|_F^2} = 1 - \frac{\mathrm{tr}(P^2)}{\|\widetilde{P}\|_F^2} \in [0, 1] \tag{3.4}$$

vanishes exactly when P is reversible and equals one exactly when no edge, self-loops included, is traversed in both directions ($P_{ij}P_{ji} = 0$ for all i, j). For a permutation kernel, κ_{circ} is the fraction of tokens lying on cycles of length at least three. Moreover, S is the operator nearest to P in $\|\cdot\|_{\mathrm{HS}(\pi)}$ among all operators self-adjoint in $L^2(\pi)$, and

$$\begin{aligned} \mathcal{I}(P) &= \sum_i \pi_i D_{\mathrm{KL}}(P_{i\cdot} \,\|\, S_{i\cdot}) = \sum_i \pi_i \,\mathrm{JS}(P_{i\cdot}, P_{i\cdot}^\dagger) \\ &= H_\pi(S) - H_\pi(P) \in [0, \min(\log 2, \sigma/2)], \end{aligned} \tag{3.5}$$

where $H_\pi(M) = -\sum_{ij} \pi_i M_{ij} \log M_{ij}$ is the stationary conditional entropy and $\sigma = D_{\mathrm{KL}}(F \,\|\, F^\top)$ is the entropy production of the stationary chain.

The proof is in Section A.4.

The circulation index therefore runs from zero, exactly for reversible heads, to one, exactly when no edge, self-loops included, is traversed in both directions; among operators self-adjoint in $L^2(\pi)$, S is the one nearest to the head in $\|\cdot\|_{\mathrm{HS}(\pi)}$; and the circulation information $\mathcal{I}(P)$ is the entropy that S has beyond P. The proof rests on two facts: symmetric and skew matrices are orthogonal, and $\langle \widetilde{P}, \widetilde{P}^\top \rangle = \sum_{ij} P_{ij} P_{ji} = \mathrm{tr}(P^2)$. Three features make the decomposition useful in practice. It is *canonical*: no basis, threshold or fitted quantity is involved. Its shares $1/\|\widetilde{P}\|_F^2$, $\|\widetilde{E}\|_F^2/\|\widetilde{P}\|_F^2$

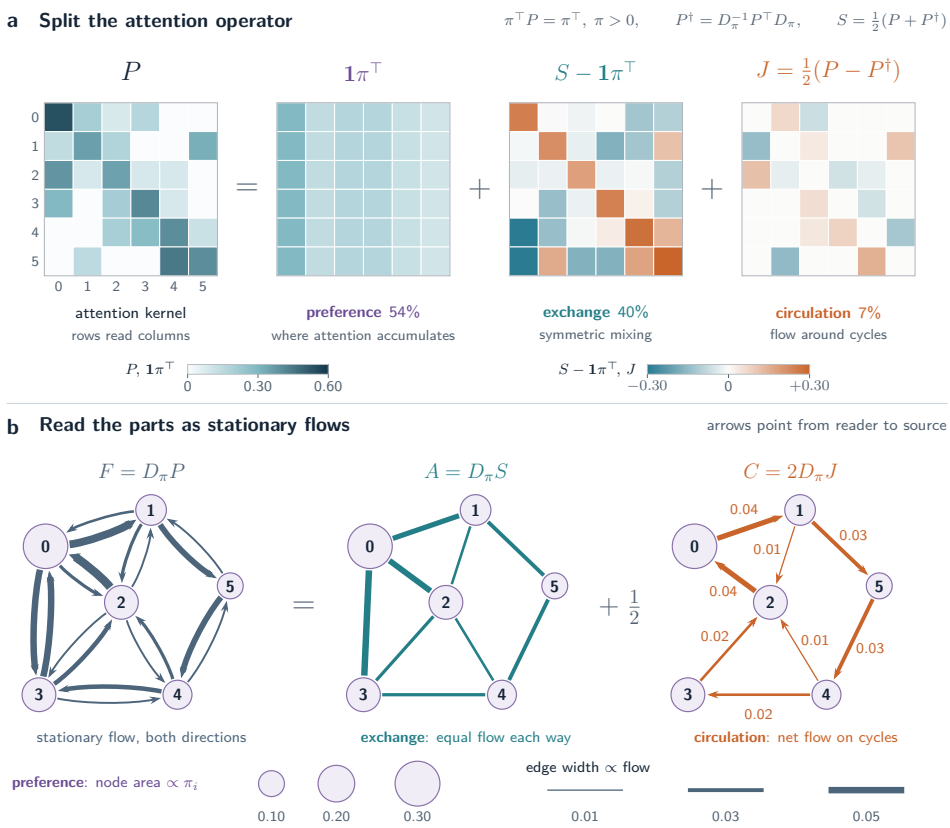


Figure 3.1: **Preference, exchange and circulation.** A constructed six-token attention kernel; Figure 3.2 decomposes a trained head. **a** The kernel is the sum of a preference $\mathbf{1}\pi^\top$, whose every row is the stationary law, a reciprocal exchange $S - \mathbf{1}\pi^\top$ and a circulation J (Theorem 3.1). The three parts are orthogonal in the head's stationary geometry, and the percentages are their shares of $\|\widetilde{P}\|_F^2$. Signed matrices are orange where positive and teal where negative. **b** The same split read as stationary flows, $F = A + \frac{1}{2}C$. Node area is the stationary mass π_i, the preference. Reciprocal traffic A moves equal amounts in both directions; the current C is divergence-free and runs around three cycles that share the edge $0 \to 1$. Tokens 0 and 3 exchange more than any other pair yet carry no current. Every positive edge is drawn, and self-loops appear only in the matrices.

and $\|\widetilde{J}\|_F^2/\|\widetilde{P}\|_F^2$ sum to one and describe how a head spends its operator energy. And the index is cheap: $\mathrm{tr}(P^2)$ is the sum of the two-step return probabilities over starting tokens and does not depend on π, so the whole calculation costs one stationary solve. Earlier work reported the ratio $x = \|\widetilde{J}\|_F/\|\widetilde{S}\|_F$. For each kernel, $\kappa_{\mathrm{circ}} = 2x^2/(1+x^2)$. The layer-6 ratios of 0.383–0.541 in the vision transformers of Section 10.3 are medians across images of means over heads, spanning two image banks. Applying the formula to these summaries gives 26–45% of maximal circulation for a kernel with that ratio; it does not give a median of the kernels' indices. The circulation information $\mathcal{I}(P)$ expresses the same content in nats: it is the stationary Jensen–Shannon divergence between a head and its own time reversal.

Figure 3.1 draws the three parts for a constructed six-token kernel, first as matrices and then as stationary flows, in which the two tokens that exchange most carry no current.

3.3 Sinks and cycles: the score-level criterion

Attention is computed from scores. Write $P = \mathrm{softmax}(s)$ row-wise, $a = s - s^\top$ for the score asymmetry and $H = I - \mathbf{1}\mathbf{1}^\top/N$ for the centering matrix.

Theorem 3.2 (Every attention asymmetry is a sink plus a cycle)**.** *Let P be softmax attention with finite scores s.*

(a) Unique split. *$a = (\mathbf{1}\psi^\top - \psi\mathbf{1}^\top) + r$ with $\psi = \frac{1}{N}a^\top\mathbf{1}$ and $r = HaH$. The two terms are orthogonal: r has zero row and column sums, and the first term is a potential difference. Hence, up to query-only terms, which cancel in softmax,*

$$P = \mathrm{softmax}(\underbrace{\tfrac{1}{2}(s + s^\top)}_{\textit{symmetric}} + \underbrace{\tfrac{1}{2}\mathbf{1}\psi^\top}_{\textit{sink}} + \underbrace{\tfrac{1}{2}r}_{\textit{cycle}}). \tag{3.6}$$

(b) Only cycles circulate. *$J = 0$ if and only if $r = 0$, that is, if and only if $P = \mathrm{softmax}(\textit{symmetric} + \textit{key bias})$. Around every cycle of tokens $i_1 \to \cdots \to i_k \to i_1$, the* cycle affinity *satisfies $\sum_t \log(P_{i_t i_{t+1}}/P_{i_{t+1} i_t}) = \sum_t r_{i_t i_{t+1}}$. Row normalizers and potentials telescope.*

(c) Any queries and keys. *If $s = \lambda QK^\top$ plus query-only and key-only terms, for arbitrary query and key features (tied or untied, rotary or not), then $r = \lambda(Q_cK_c^\top - K_cQ_c^\top)$ with $Q_c = HQ$ and $K_c = HK$. For a shared input X and $s = XWX^\top$, this reads $r = 2X_cW_{\mathrm{sk}}X_c^\top$ with $X_c = HX$ and $W_{\mathrm{sk}} = \frac{1}{2}(W - W^\top)$. Circulation therefore requires the skew form to be nonzero on the span of the centered tokens, and collinear centered tokens cannot circulate.*

(d) Entropy production is a score–current pairing. *$\sigma = D_{\mathrm{KL}}(F\|F^\top) = \langle C, s\rangle_F = \frac{1}{2}\langle C, r\rangle_F$; for dot-product heads, $\sigma = \lambda\langle Q, CK\rangle_F$.*

The proof is in Section A.4. The theorem separates two things that are routinely conflated. A key bias, the mechanism by which attention sinks are commonly described (Gu et al., 2025), adds only a potential. It can move stationary mass, reshape reciprocal traffic and rescale an existing current, but it can never create circulation. Circulation requires a cycle in the scores. Part (c) gives a solver-free screen: $\|r\|_F^2 = 2\lambda^2[\mathrm{tr}(Q_c^\top Q_c K_c^\top K_c) - \mathrm{tr}((Q_c^\top K_c)^2)]$ costs $O(Nd^2)$ from the (possibly rotated) queries and keys. The criterion concerns the scores a model actually computes on its actual tokens. For scores of the shared bilinear form $s = XW_{QK}X^\top$, a nonzero skew part of W_{QK} is necessary for circulation but not sufficient (Saponati et al., 2025; O'Desky, 2026).

Figure 3.2 shows both splits on a previous-token head of BERT.

Part (c) makes Theorem 3.2 a design criterion, and Table 3.1 applies it to common choices. Two rows deserve comment. Tying queries to keys removes circulation in every head, on every input, and at every layer. But tying also makes the scores a Gram matrix, and a Gram head cannot, on geometric average around any cycle, attend to a neighbor more than to the reading token itself (Section 10.1). Shared projections with a diagonal rescaling can be symmetric without being positive semidefinite, and can escape that restriction. Section 10.5 measures what this difference costs a circulation-free network that must compute direction. And a causal decoder has no stationary circulation at all. If P is lower triangular (reader i sees $j \le i$) with $P_{11} = 1$ and $P_{ii} < 1$ for $i > 1$, the stationary equations, solved from the last token upwards, force $\pi = e_1$, so $F = e_1e_1^\top$ and $C \equiv 0$; a nonzero skew current supported on allowed edges would need both (i, j) and (j, i) to be allowed, which a strict causal mask permits only for $i = j$. Its asymmetry is pure hierarchy, and the full-token construction is vacuous rather than merely uncertified. Bidirectional sub-blocks, such as prefix-LM prefixes, joint text–image blocks (Esser et al., 2024) and the attention of masked diffusion language

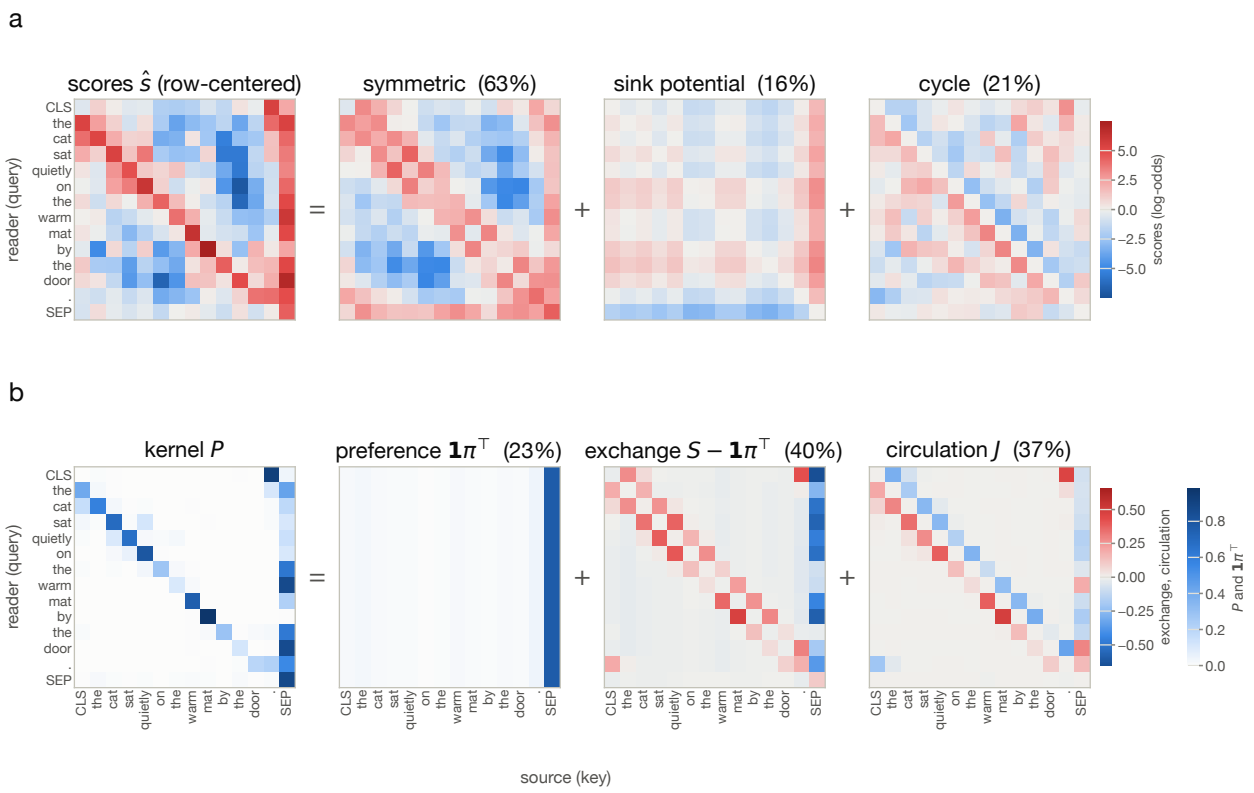


Figure 3.2: **A trained head, decomposed two ways.** BERT-base, layer 6, head 11 (a previous-token head in the sense of Clark et al., 2019) on one sentence. **a** Scores, with query-only terms removed ($\hat{s} = (\log P)H$), split orthogonally into a symmetric part, a key-bias *sink* potential and a doubly centered *cycle* (Theorem 3.2); percentages are shares of $\|\hat{s}\|_F^2$. **b** The kernel split into preference, exchange and circulation (Theorem 3.1); percentages are shares of $\|\widetilde{P}\|_F^2$ in stationary coordinates. The preference is the [SEP] sink. The circulation is the directed band of previous-token reading, and its cycle closes through the special tokens ([CLS] reads the final period). The cycle part of the scores vanishes exactly when the circulation does. Reader-to-source orientation; red is positive.

models (Nie et al., 2025; Sahoo et al., 2024), are eligible. The rotary rows follow from the cycle affinity of rotary scores around a triangle of tokens. Untied rotary attention can therefore circulate even when content is constant. For constant content and fixed token offsets, low frequencies act almost like a potential. Barbero et al. (2025b) find that positional heads use middle and high frequencies.

3.4 Exchange smooths and circulation transports

The generator $P - I = (S - I) + J$ separates a self-adjoint, dissipative part from a skew-adjoint, conservative part. In the language of transport, exchange diffuses and circulation advects.

Proposition 3.3 (Generator view). *(a)* $\langle f, Jf\rangle_\pi = 0$ *for all* f*, so the Dirichlet form* $\langle f, (I - P)f\rangle_\pi = \frac{1}{2}\sum_{ij} A_{ij}(f_i - f_j)^2$ *depends on* S *alone.*

(b) For every real α*,* $\mathrm{Re}\, W(\widetilde{S} + \alpha\widetilde{J}) = W(\widetilde{S})$*, where* W *is the numerical range. Every nontrivial*

Table 3.1: **Which designs can circulate?** Consequences of Theorem 3.2 for common score and normalization choices, assuming finite scores on full support. "None" means $J = 0$ identically. "Possible" means the cycle term is generically nonzero. Every row is derived, and checked numerically, in Bruce (2026g, Appendix B).

Design	Circulation	Reason
Symmetric scores; tied $Q = K$; tied L_2 attention (Kim et al., 2021)	none	$r = 0$; $\pi_i \propto \sum_j e^{s_{ij}}$
Shared projection with diagonal rescaling, $Q = XWD_q$, $K = XWD_k$ (Kowsher et al., 2025)	none	$QK^\top$ symmetric (possibly indefinite)
Query or key biases (sinks as key biases)	presence unchanged	potential terms only
Even relative bias $b(\lvert i - j\rvert)$ on symmetric content scores, e.g. ALiBi's bias (Press et al., 2022)	none	no antisymmetric part
Odd bias linear in the offset, on symmetric content scores	none	a potential on a line
Nonlinear odd relative bias (T5 buckets, learned tables; Raffel et al., 2020; Liu et al., 2021)	possible	odd part is not a potential
RoPE with tied queries and keys (Su et al., 2021)	none	rotated Gram matrix is symmetric
RoPE with untied queries and keys	possible, even for constant content	$r = \lambda(Q_c K_c^\top - K_c Q_c^\top)$ after rotation
Row-dependent temperatures $s_{ij} = \tau_i c_{ij}$, c symmetric (shared-QK with normalized keys, Kitaev et al., 2020)	possible	$(\tau_i - \tau_j)c_{ij}$ need not be a potential
Doubly stochastic (Sinkhorn) attention (Sander et al., 2022)	$J = \frac{1}{2}(P - P^\top)$	π uniform, so $P^\dagger = P^\top$
GATv1 scores LeakyReLU$(u_i + v_j)$ (Veličković et al., 2018)	none on one linear piece	only the kink breaks the potential
GATv2 (Brody et al., 2022); graph attention on graphs with cycles	possible	
Any positive weights on a tree-structured support	none	detailed balance solvable edge by edge
Causal (lower-triangular) mask	$C \equiv 0$	stationary flow sits on the first token; no nonzero feasible current exists

eigenvalue of $S + \alpha J$ therefore has real part in $[\lambda_{\min}(\widetilde{S}),\ 1 - \mathrm{gap}(S)]$, and the continuous-time semigroup $e^{t(S+\alpha J - I)}$ contracts mean-zero functions at rate at least $\mathrm{gap}(S)$. Circulation never slows relaxation, the finite-state analogue of the accelerating effect of divergence-free drift (Hwang et al., 2005).

(c) For content X with π-centered part X_0, averaging over orientations gives $\frac{1}{2}(\|P_a X_0\|_\pi^2 + \|P_{-a} X_0\|_\pi^2) = \|SX_0\|_\pi^2 + a^2\|JX\|_\pi^2$, where $P_\alpha = S + \alpha J$. Averaged over the two orientations, circulation preserves more token variance than stilling does. A single orientation can go either way, through $\langle SX_0, JX_0\rangle_\pi = \frac{1}{2}\operatorname{tr}(\widetilde{X}_0^\top[\widetilde{S}, \widetilde{J}]\widetilde{X}_0)$, which vanishes for all content exactly when $\widetilde{P}$ is normal.

(d) For a circulant head, $P_{ij} = k(j - i)$ on a periodic domain, π is uniform. S and J carry the

even and odd parts of k, so stilling is zero-phase filtering and reversal conjugates the phase. For a unit shift, S averages the two neighbors (a blur) and J is the central difference (a derivative).

The proof is in Section A.4. Part (d) is the cleanest picture of the decomposition: in a convolution-like head (Cordonnier et al., 2020), exchange is a blur and circulation is a derivative. Self-attention is often described as a low-pass filter (Wang et al., 2022). In this language, exchange is the zero-phase low-pass component and circulation carries the phase. The picture is exact on periodic domains. Boundaries and the class token make π nonuniform in real vision transformers, which is why the stationary weighting matters. By part (c), exchange and circulation outputs are π-orthogonal for every content when a head is normal, for instance circulant. DeiT's attention maps visually resemble block-circulant matrices (Han et al., 2026). Consistent with near-normality, in the vision transformers of Section 10.4 the delivered circulation B_J (Section 10.2) is nearly orthogonal to the delivered reciprocal output $\sum_h S_h V_h W_O^{(h)}$. At layer 6, median cosines lie between -0.02 and $+0.14$, and 98–99.95% of the energy of B_J is orthogonal to the reciprocal output. This is a Euclidean observation after the output projection, not an instance of the proposition.

3.5 Parity: quadratic readouts cannot see direction

The sweep $P_t = S + tJ$ of (1.3) runs from $P^\dagger$ at $t = -1$ through S to P at $t = 1$, and every P_t is a walk with stationary distribution π and flow ratio $|t|\Phi$. The key observation is that $P_{-t} = P_t^\dagger$: running the sweep backwards is reversing time.

Theorem 3.4 (Parity)**.** *Let P be row-stochastic with stationary $\pi > 0$, and let f be a normalized amplifier with $\mathbb{E}[K^2] < \infty$. Write v_k, μ_k for the eigenpairs of S and $J_{kl} = \langle v_k, Jv_l\rangle_\pi$.*

(i) Quadratic readouts are even. *For every real Y, the function $F(t) = \langle Y, f(P_t)Y\rangle_\pi$ satisfies $F(t) = F(-t)$: all its odd-order terms vanish. The spectrum of P_t is likewise the same set at t and $-t$.*

(ii) Their size. $|F(t) - F(0)| \le \frac{1}{2}\, t^2\Phi^2\, \mathbb{E}[K(K-1)]\, \|Y\|_\pi^2$.

(iii) Cross readouts and eigenvectors pay first order. *For real X, Y, $G(t) = \langle X, f(P_t)Y\rangle_\pi$ has $G'(0) = \langle X, \mathcal{D}_f Y\rangle_\pi$, where $\mathcal{D}_f$ has modal entries $\langle v_k, \mathcal{D}_f v_l\rangle_\pi = f[\mu_k, \mu_l]\, J_{kl}$ with the divided difference $f[\mu_k, \mu_l] = (f(\mu_k) - f(\mu_l))/(\mu_k - \mu_l)$ (and $f'(\mu_k)$ when the two coincide). This array is antisymmetric, so $G'(0)$ changes sign when X and Y are exchanged. If μ_j is a simple eigenvalue of S, the corresponding eigenvalue of P_t has zero derivative at $t = 0$, while its eigenvector, normalized by $\langle v_j, v_j(t)\rangle_\pi = 1$, moves at first order: $v_j(t) = v_j + t \sum_{l \ne j} \frac{J_{lj}}{\mu_j - \mu_l}\, v_l + O(t^2)$.*

(iv) Output energies are not quadratic readouts. $\frac{d}{dt}\|P_t Y\|_\pi^2\big|_{t=0} = \langle Y, [S, J]\, Y\rangle_\pi$, *which vanishes for every Y if and only if P commutes with its time reversal ($PP^\dagger = P^\dagger P$).*

Proof of (i). $P_{-t} = S - tJ = P_t^\dagger$. The adjoint reverses products and fixes real scalars, so $f(P_t)^\dagger = f(P_t^\dagger) = f(P_{-t})$. For a real vector and any real matrix M, $\langle Y, MY\rangle_\pi = \langle M^\dagger Y, Y\rangle_\pi = \langle Y, M^\dagger Y\rangle_\pi$, because the real π-inner product is symmetric. Hence $F(t) = \langle Y, f(P_t)^\dagger Y\rangle_\pi = F(-t)$. For the spectrum, $\mathrm{spec}(M^\dagger)$ is the complex conjugate of $\mathrm{spec}(M)$, and the spectrum of a real matrix is closed under conjugation. □

Parts (ii)–(iv) are proved in Section A.4. The first-order correction to the amplified operator is skew-adjoint, so every real quadratic readout cancels it. Cross readouts and eigenvectors can respond at first order, while simple eigenvalues have zero first derivative. The divided difference

Table 3.2: **Which statistics can see the direction of the flow.** Along the sweep $P_t = S + tJ$ a statistic unchanged by time reversal is even in t; smooth ones then move at second order, and norms of odd objects at first order in $|t|$, as a synthetic check shows.

Statistic	**Under $P \mapsto P^\dagger$**	**Change at small flow**
Anything computed from S alone (its spectrum, gap, eigenvectors)	unchanged	none
Spectrum of P as a set; canonical singular values; gap of $P^\dagger P$	unchanged	second order (simple values)
Entropy production; circulation index and circulation information (Theorem 3.1)	unchanged	second order
Quadratic readouts $\langle Y,\, f(P)Y\rangle_\pi$ (Theorem 3.4)	unchanged	second order
Flow ratio Φ; commutator norm $\|PP^\dagger - P^\dagger P\|_\pi$; height of the numerical range	unchanged	first order in $\|t\|$
Cross readouts $\langle X,\, f(P)Y\rangle_\pi$; eigenvectors, projectors	odd part	first order, signed
Output energies $\|f(P)Y\|_\pi^2$ when $PP^\dagger \neq P^\dagger P$	changed	first order, signed

decides which side a readout lands on: it is symmetric in (μ_k, μ_l) and the flow array is antisymmetric, so a symmetric read-out array cancels them and an antisymmetric one does not.

Part (i) of Theorem 3.4 covers every readout that cannot tell P from $P^\dagger$: quadratic forms $\langle Y,\, f(P)Y\rangle_\pi$, the spectrum as a set, simple eigenvalues, the collapse ratio below, and suppression factors. These are the collapse-type readouts. A degenerate eigenvalue of S can split at first order into a conjugate pair; the set is unchanged by reversing the flow, but can move at first order from the reversible spectrum: on a plane where S acts as λI and J as a rotation of strength c, the eigenvalues of P_t are $\lambda \pm ict$ (on the lazy 4-cycle, $\frac{1}{2} \pm it/2$, a machine-precision check). The Dobrushin leak and the ℓ_∞ distance of Section 2.5, for P with $\pi > 0$, are not π-quadratic readouts, and they are not even in the flow. The orientation-type readouts are those that can tell time's direction: cross terms $\langle X,\, f(P)Y\rangle_\pi$ that carry content from one mode to another, eigenvectors and spectral projectors, and output energies $\|f(P)Y\|_\pi^2$ of operators that do not commute with their time reversal. Subspace angles measure the unsigned size of a turn and can respond at first order in $|t|$. Evenness gives second order only for smooth readouts; Proposition 3.7 shows when the distance to collapse, a norm, is flat to first order. And the constant in (ii) is informative only when $\frac{1}{2}\Phi^2\mathbb{E}[K(K-1)]$ is small compared with the readout; the parity and bound are exact throughout the sweep, but the bound need not be tight. The same symmetry sorts the statistics used to detect directed structure (Table 3.2). A statistic that is unchanged by $P \mapsto P^\dagger$ is an even function of the flow along the sweep, so it cannot tell a circulation from its reverse at any order: it measures how much circulation there is, never its sign. Such statistics move at second order when they are smooth in t; norms of odd objects, such as the flow ratio itself, move at first order in $|t|$. Only statistics that change under reversal can see the sign. Cross readouts and eigenvectors can see it at first order, as a synthetic check shows.

Corollary 3.5 (The collapse ratio travels)**.** *For a directed operator let v_C be a π-unit eigenvector of S at the content eigenvalue μ_C, and read the collapse ratio through $\kappa_{\mathrm{col}}(P)^{-1} := \langle v_C,\, f(P)v_C\rangle_\pi$, which reduces to Definition 2.7 when P is reversible. Then $\kappa_{\mathrm{col}}(P_t)^{-1} = \kappa_{\mathrm{col}}(P_{-t})^{-1}$ exactly, and $|\kappa_{\mathrm{col}}(P)^{-1} - \kappa_{\mathrm{col}}(S)^{-1}| \le \frac{1}{2}\Phi^2\,\mathbb{E}[K(K-1)]$.*

Proof. Apply Theorem 3.4(i)–(ii) to f with $Y = v_C$, for which $F(0) = f(\mu_C) = \kappa_{\mathrm{col}}(S)^{-1}$. □

For the four species, $\mathbb{E}[K(K-1)]$ is $L(L-1)$ for depth, $2\alpha^2/(1-\alpha)^2$ for the restart, $L(L-1)b^2$ for the residual stack and σ^2 for the residual stream. The collapse law of Section 2.3 therefore carries

over to one directed operator with an error that is even in the flow, second order, and explicitly bounded. (For a directed operator $\langle v_C,\, f(P)v_C\rangle_\pi$ can be small or negative when the flow is large, so we state the corollary for $\kappa_{\mathrm{col}}^{-1}$.)

A network applies different operators at different layers, and the symmetry becomes a statement about the order in which they are applied.

Proposition 3.6 (Palindromes are even; other words pay first order)**.** *Let $P^{(1)}, \dots, P^{(L)}$ share the stationary distribution $\pi > 0$, with parts $S^{(\ell)}, J^{(\ell)}$ and flow ratios Φ_ℓ, and let $W(t) = P_t^{(L)} \cdots P_t^{(1)}$. For real Y, $F_W(t) = \langle Y,\, W(t)Y\rangle_\pi$ equals the same readout of the reversed word at $-t$. Hence:*

(i) *if the word is a palindrome ($P^{(\ell)} = P^{(L+1-\ell)}$), F_W is even in t; powers of one operator are the simplest palindromes;*

(ii) *for any word, $F_W'(0) = \sum_\ell \langle u_\ell,\, J^{(\ell)} w_\ell\rangle_\pi$ with $u_\ell = \big(S^{(L)} \cdots S^{(\ell+1)}\big)^\dagger Y$ and $w_\ell = S^{(\ell-1)} \cdots S^{(1)} Y$, and $|F_W'(0)| \le \sum_\ell \Phi_\ell \, \|u_\ell - w_\ell\|_\pi \, \|Y\|_\pi$.*

The proof is in Section A.4. This proposition sets the scope of the parity statement. A stack of distinct layers is not a palindrome, and its collapse-type readouts generally pay at first order in the flow; the first-order term is a ledger computable in one forward and one backward sweep, charging each layer its flow times the mismatch between what the layers before it and the layers after it do to the readout. The shared stationary distribution is essential; with layer-specific distributions the reversible parts are not self-adjoint in one geometry. And second order is a floor rather than an artefact of the bound: interleaving a layer with its own time reversal, as in the palindrome $P\,P^\dagger P$, still pays at second order, as a machine-precision check shows.

Corollary 3.5 says the collapse ratio of one operator barely notices the flow. The distance of the amplified operator from its rank-one limit, $d_f(P) = \|f(P) - \mathbf{1}\pi^\top\|_\pi$, is a norm rather than a quadratic form, and needs its own statement.

Proposition 3.7 (Collapse distances off the reversible manifold)**.** *Let P be irreducible with stationary $\pi > 0$, let f be a normalized amplifier with $\mathbb{E}[K^2] < \infty$, and let $\bar\gamma = 1 - \mu_2$ be the spectral gap of S.*

(i) Order. *Suppose $d_f(S) > 0$ and f takes one and the same value at every content eigenvalue μ_k of S where $|f(\mu_k)|$ is largest, which holds whenever P is lazy. Then $d_f(P_t) = d_f(S) + O(t^2)$. Without the condition d_f can move at first order.*

(ii) Half-clock law. *If P is lazy, $d_f(P) \le \mathbb{E}\big[(1-\bar\gamma)^{K/2}\big]$, with equality for the lazy cycle at fixed depth: circulation can at most halve the exponential collapse rate of a lazy operator.*

(iii) Operators that commute with their time reversal. *If $PP^\dagger = P^\dagger P$, then $d_f(P) = \max\{|f(\zeta)| : \zeta \in \mathrm{spec}(P),\ \zeta \ne 1\}$; each eigenvalue has the form $\zeta = \xi + i\eta$ with $\xi \in \mathrm{spec}(S)$ and $|\eta| \le \Phi$; and for $\xi \in (0,1)$*

$$|f(\xi + i\eta)|^2 \;=\; f(\xi)^2\Big[1 + \frac{\eta^2}{\xi^2}\big(\mathbb{E}_\xi K - \mathrm{Var}_\xi\, K\big)\Big] + O(\eta^4),$$

where $\mathbb{E}_\xi$, Var_ξ refer to the tilted lottery $\Pr_\xi[K=k] \propto p_k \xi^k$. For the Poisson lottery, $|f(\xi+i\eta)| = f(\xi)$ exactly.

(iv) Mixed-Poisson lotteries are never slowed. *If K is Poisson with a random mean (this includes the residual stream and the restart), then $d_f(P) \le d_f(S)$ for every irreducible P. Depth admits no such comparison in either direction.*

The proofs are in Section A.4. Part (i) completes the parity statement for one operator: for lazy operators the distance to collapse, not only the quadratic readouts, is flat to first order in the flow.

The condition is needed. A non-lazy operator with eigenvalues $\{1, 0.2, -0.2\}$ plus a three-cycle flow has, for one step ($f(z) = z$), $d_f = 0.2 + \Phi$ exactly (a machine-precision check), because the gains $f(\pm 0.2) = \pm 0.2$ have equal size and opposite signs and their singular values cross; in a synthetic check, random lazy operators, including ones whose top content eigenvalue is degenerate, show slope 2.000. Part (ii) rests on the classical fact that a lazy operator satisfies $P^\dagger P \preceq S$ in the π-geometry (Mihail, 1989; Fill, 1991): a single step of P shrinks content by a factor of at most $\sqrt{1-\bar{\gamma}}$ whatever the flow. It is a collapse law for every lazy directed operator with constant 1 and no perturbative assumption. Part (iii) is the dispersion law of Proposition 2.10 seen from the complex plane: at a fixed real part and to second order in the imaginary part, circulation slows collapse when the tilted lottery is under-dispersed (depth: $|\zeta|^L > \xi^L$) and speeds it when it is over-dispersed (restart). For the Poisson lottery the gain is unchanged exactly. For depth, restarts and residual streams the tilted lottery is again deterministic, geometric or Poisson, so the sign is that of $\mathbb{E}K - \operatorname{Var} K$ in Table 1.1; for other lotteries tilting can reverse it (in a machine-precision check, a lottery on $\{0, 10\}$ with probabilities $0.01, 0.99$ is under-dispersed, yet circulation speeds its collapse at $\xi = 0.5$). Part (iv) covers the amplifiers that are continuous-time walks (mixed-Poisson lotteries), for which the classical lesson is that non-reversibility cannot slow the worst-case rate (Hwang et al., 1993). In discrete time it can speed mixing (Chen et al., 1999; Diaconis et al., 2000) but can also slow it: for depth the lazy cycle has $d_f(P) > d_f(S)$, while a synthetic check finds random lazy operators with ratio 0.67. Off the commuting class the spectrum gives only a floor: operators with the same π and the same spectrum can have different distances (Proposition A.3).

Three further statements are collected with their proofs in Section A.4: the flow ratio as the half-height of the numerical range (Proposition A.2), which Section 1.2 uses; the spectrum as only a floor for the collapse distance off the commuting class, with a pair of operators that share their spectrum (Proposition A.3); and a quantitative bound for the eigenvector turns of Theorem 3.4(iii) (Proposition A.4).

3.6 Measured: order at small flow and size at native flow

The order is measured first, at small flow. Sixteen public checkpoints (seven autoregressive: GPT-2, Pythia, OPT; four bidirectional: BERT, RoBERTa, DistilBERT; five ViT/DeiT classifiers) give 270 per-layer operators (the first 24 of GPT-2 large's 36 layers), each averaged over heads and over the sixteen inputs of its capture and swept along $P_t = S + tJ$ for $t \in [0.003, 0.2]$ with a restart at $\alpha = 0.8$. The quadratic readout of the nine non-autoregressive checkpoints, whose operators have a positive stationary distribution, sits at slope ≈ 2.000; pooled over all 270 operators the slopes are 1.905 ($[1.885, 1.924]$) for the quadratic readout and 1.001 ($[1.001, 1.002]$) for the sine of the angle between the spectral-projector ranges, all measured on real operators along an artificial perturbation (Bruce, 2026j). The pooled values are means of per-operator slopes; brackets give descriptive 95% bootstrap intervals from resampling those operators. Both miss the bands registered for the experiment, which were copied from synthetic slopes rather than derived from the dose window (the ledger in Appendix B; the derivation is in Bruce, 2026i, Appendix E); the pooled quadratic-readout slope is pulled below two by the autoregressive checkpoints alone. The seven autoregressive checkpoints are outside the theorem's scope: their operators have no positive stationary distribution (Proposition 2.15) and were analysed using a floored positive reference in its place. Their softened slopes (1.74–1.85) reflect either the conditioning of that regularization or higher-order terms at their larger flow; the companion record (Bruce, 2026j) does not separate the two, and Theorem 3.4 itself is exact for any positive stationary π. Because the leading terms of Theorem 3.4 are exact, these slopes validate the order and the implementation on real attention;

they are not an independent test, and they say nothing about the size of the effect at the operators' own flow.

The size is measured on trained models at $t = 1$, on the operators' own flow. We use BERT-base (Devlin et al., 2019) on the sixteen WikiText-2 paragraphs of Section 2.7 (first 64 tokens) and DeiT-small (Touvron et al., 2021) on eight 224×224 crops of two natural images, head-average each layer, and use the lazy form $\frac{1}{2}(I + P)$, the residual block with $b = \frac{1}{2}$: 192 and 96 operators with median flow ratio 0.061 and 0.059 (maximum 0.14). Quadratic readouts are even to rounding ($|F(1) - F(-1)| \le 1.3 \times 10^{-15}$). Here $C = 2$; the eigenvector of P compared with v_2 is the one whose eigenvalue is nearest μ_2. The collapse-type readout $|\langle v_C, f(P)v_C\rangle_\pi - f(\mu_C)|/f(\mu_C)$ has median 0.38% / 0.29% (BERT / DeiT) for the restart at $\alpha = 0.8$, 0.27% / 0.17% for the residual stream at $\sigma = 2$, and 1.5% / 1.1% for depth 4 (maximum 5.9%), always within the bound of Corollary 3.5, which is loose on these operators (median bound 30 times the change for the restart, 4–6 times for the others). The eigenvector at the same mode turns by a median sin = 0.29 / 0.28, about 17° / 16°, and by up to 67°–74° (Figure 3.3a). The turns are large because the readout eigenvalue's neighbours are close (median local gap $\gamma_C = \min_{l \neq C} |\mu_C - \mu_l|$ of 0.032 / 0.049, comparable to the flow). Most operators therefore lie outside the window $\Phi < \gamma_C/2$ in which the projector bound of Proposition A.4 certifies perturbation theory (only 1 of 192 BERT and 18 of 96 DeiT operators lie inside), yet the first-order formula of Theorem 3.4(iii) still predicts the turns: measured over predicted has median 1.07 / 1.06 (1.02 on the one BERT operator and a median of 1.01 on the 18 DeiT operators inside the window). The one-sided law of Proposition 3.7(iv) holds on every operator (restart and residual stream: $d_f(P)/d_f(S) < 1$, maximum 0.9999993). For depth the median ratio is 1.0012 / 1.0013 (range 0.9994–1.0054 and 1.00002–1.011), in the direction of the tilted sign rule of Proposition 3.7(iii); these operators do not commute with their time reversal, so this is consistency with the rule, not a prediction of it.

To test Proposition 3.6 on real layers we balance the BERT layers of eight of the sixteen inputs to doubly stochastic form, so that they share the uniform stationary distribution (Sinkhorn, 1964; Sander et al., 2022), and sweep five-layer words drawn from three distinct layers of one input (median flow ratio of the balanced layers, not in lazy form, 0.13). Forty-eight palindromes (*abcba*) have slope 2.000 (range 1.99995–2.0001) and a relative reversal difference $|F_W(1) - F_W(-1)|/|F_W(0)|$ below 10^{-15} at native flow; the 48 matched non-palindromes (*abcab*) have slope 1.000 (range 0.974–1.015) and a median relative reversal difference of 7.6%, all measured on real operators along an artificial perturbation (Figure 3.3b). Real stacks are not palindromes, so this is the scope limit of the parity statement, measured.

At the causal end π degenerates and Φ is undefined (Proposition 2.15). The Frobenius fingerprints used to compare model families, which set the antisymmetric part of a diagonally rescaled operator against its symmetric part, are large for causal families when measured on trained models (Bruce, 2026j), and triangularity forces them: for any lower-triangular operator they are a closed function of the squared diagonal and off-diagonal masses, so they do not measure learned directedness.

Lemma 3.8 (The causal fingerprint is fixed by triangularity)**.** *Let P be strictly causal and W any positive diagonal matrix; set $M = WPW^{-1}$ and $M_\pm = \frac{1}{2}(M \pm M^\top)$. Then $\|M_-\|_F = \|M_+^{\mathrm{off}}\|_F$, where M_+^{off} is M_+ with its diagonal removed, and*

$$\Phi_{\mathrm{F}}^2 := \frac{\|M_-\|_F^2}{\|M_+\|_F^2} = \frac{m_{\mathrm{off}}}{m_{\mathrm{off}} + m_{\mathrm{self}}}, \qquad m_{\mathrm{off}} = \tfrac{1}{2}\sum_{i>j} M_{ij}^2, \quad m_{\mathrm{self}} = \sum_i P_{ii}^2.$$

The proof is in Section A.4. The fingerprint of a causal operator is thus a closed function of its squared diagonal mass and its squared below-diagonal mass. The numerical examples and captured-layer summaries below use $W = I$ unless another convention is stated. It depends on how

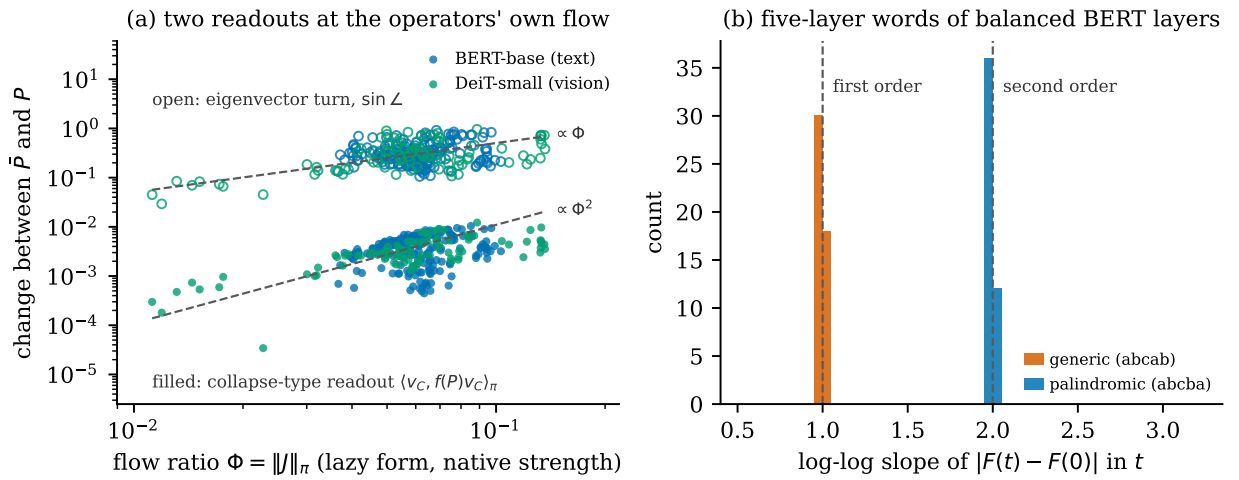


Figure 3.3: **The parity theorem at the operators' own flow**. (a) Measured on trained models: for every lazy, head-averaged per-layer operator of BERT-base and DeiT-small, the relative change of the collapse-type readout $\langle v_C,\, f(P)v_C\rangle_\pi$ between S and P (filled; restart at $\alpha = 0.8$) together with the sine of the angle through which the corresponding eigenvector turns (open), both against the flow ratio; dashed guides have slopes 2 and 1. (b) On real operators along an artificial perturbation: log-log slopes of $|F_W(t) - F_W(0)|$ for five-layer words of balanced BERT layers: palindromes are exactly second order, their non-palindromic rearrangements first order. The panels write $\bar{P}$ for the reversible part, which the book writes S.

the off-diagonal mass is spread (in a machine-precision check, with the same mask and diagonal, a last row $(0.5, 0, 0.5)$ gives $\Phi_{\mathrm{F}}^2 = 1/7$ and $(0.25, 0.25, 0.5)$ gives $1/9$), but it carries no information about learned directedness beyond triangularity. In a machine-precision check (Bruce, 2026j), the identity reconstructs the fingerprint of 156 captured causal layers to 4.9×10^{-16}, and fails for bidirectional operators, whose supports overlap (median identity residual $|\|M_-\|_F - \|M_+^{\mathrm{off}}\|_F| / \|M_+^{\mathrm{off}}\|_F$ of 0.114 on 54 encoder layers). Comparisons of fingerprints must state both the convention and the aggregation. Measured on trained models, the ratio of summed masses $\sum m_{\mathrm{off}} / \sum m_{\mathrm{self}} = 7.89$ over those 156 layers (mean of per-layer ratios 9.82) corresponds to $\Phi_{\mathrm{F}} = 0.942$ at that pooled ratio; on GPT-2 small, Φ_{F} at the mean of the per-layer ratios (8.82) is 0.948, while the mean of the 192 per-layer values is 0.877 (median 0.954), because $r \mapsto \sqrt{r/(1+r)}$ is concave; and the causal-family mean 0.860 (descriptive 95% bootstrap interval $[0.842, 0.879]$ over seven checkpoints) of Bruce (2026j) was computed by averaging the per-(layer, head) ratios within each checkpoint, then weighting checkpoints equally. Those ratios use plain Frobenius norms of the additive split formed with a floored positive reference. Finally, Φ_{F} is not the flow ratio: the two are not ordered in general (in a machine-precision check, padding the lopsided triangle at $s = 0.1$ with 30 identity tokens leaves $\Phi = 0.173$ while Φ_{F} falls from 0.229 to 0.044; the padded operator is reducible, and π is taken uniform).

3.7 What holds in trained networks

Exact for every network of the stated class. The split of a head into preference, exchange and circulation (Section 3.2) is an identity for every Markov kernel with a positive stationary law; for every unmasked softmax head with finite scores, its score asymmetry also splits into a sink and a

cycle (Section 3.3), at every layer and input. Causal heads have no positive stationary law.

Exact under premises checked network by network. The parity law (Section 3.5, measured in Section 3.6): exact, for stationary-weighted quadratic readouts, for one amplified operator and for palindromic words of layers that share a positive stationary law. A stack of distinct layers is not a palindrome, and its readouts generally pay at first order (measured median log-log slope 1.000 in the circulation strength, against 2.000, for 48 matched pairs of non-palindromic and palindromic five-layer words of balanced BERT layers from eight inputs).

The native-flow measurements of Section 3.6 average heads, which assumes that the heads of a layer share their value and output maps; they use the lazy form, which halves the flow and the gaps; and they rest on 64-token texts, eight crops of two images and two small encoders. The stationary distributions of the encoder operators are concentrated (median $\pi_{\max}/\pi_{\min}$ of 391 and 378; the median effective support $1/\sum_i \pi_i^2$ is 6% and 13% of the tokens), so readouts in the π-geometry weight a few tokens heavily. The residual share b is assumed in these measurements, not measured from the trained networks. Much of the evidence beyond the checks released with the theory (Bruce, 2026i) rests on the companion record (Bruce, 2026j), and several quoted results come from experiments whose registered outcome was mixed or negative; the evidence ledger (Appendix B) marks them.

For the split and for the measurements of circulation in trained networks, the claims, their evidence and their limits are tabulated once, with the closing of the chapter on the arrow (Section 10.9).

Chapter 4

One theory across graphs, denoisers and context mixing

The operator of Chapter 1 is not peculiar to attention. Chapter 0 observes that the same object appears in graph propagation, label propagation, adaptive denoising and test-time mixing, and that the same laws govern it there (Section 0.3.1). This chapter takes the laws of Chapters 1–3 to those settings: they apply wherever a row-stochastic kernel is amplified, by depth, by a restart or by a residual stream, on the premises stated with them. Section 4.1 reads graph and label propagation through the collapse law; Section 4.2 shows that iterative denoisers, deep equilibrium models and test-time context mixing are instances of the anchored iteration of Section 1.4; Section 4.3 asks where gain should sit relative to a second map, and Section 4.4 when propagation helps frozen features at all. Section 4.6 says what of this holds in trained networks.

4.1 Graphs and label propagation

On a graph the operator is the normalized adjacency of (1.1), and the lotteries of Table 1.1 are familiar propagation schemes: depth is graph convolution, and the restart is label spreading, APPNP's teleport and GCNII's initial residual (Section 1.3). Figure 4.1 evaluates Table 1.1 at the content gap of the lazy propagation operator on a 1,500-node connected subgraph of Cora ($\beta = 0.0261$): depth collapses, a constant-share residual stack collapses more slowly, and the restart and the bounded-total-update stream stay bounded. The realized graph-convolution and APPNP/GCNII modal gains on that subgraph match the corresponding closed forms on the probed eigenmodes with maximum absolute error 1.5×10^{-14}, and the collapse-ratio identity holds with relative error 6.5×10^{-16}, in machine-precision checks (Bruce, 2026j). In controls on Cora and a synthetic two-factor testbed, with one seed each, four-layer perceptrons share no mixing operator across samples and show no cross-sample collapse of this kind, as expected if the phenomenon belongs to mixing operators and not to depth as such. Depth four alone met the convergence check; the tested depths from 8 to 64 missed it, although the eight-layer synthetic model still reduced its loss and performed above chance. The deepest models' final-layer features concentrated (cosine up to 1).

The path-law form of the same laws is measured on graph spectra and on propagation accuracy in Section 2.7 (Bruce, 2026k); those measurements are not repeated here.

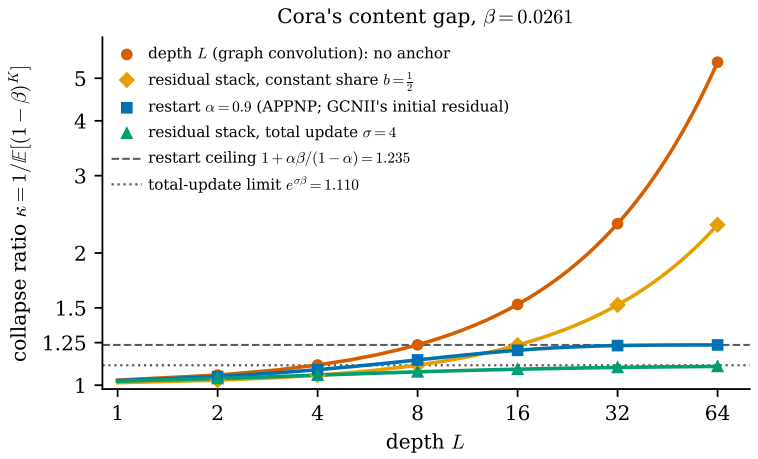


Figure 4.1: **Illustration: the collapse ratio at Cora's content gap** ($\beta = 0.0261$, measured on a 1,500-node connected subgraph of Cora). Closed forms of Table 1.1: depth (no anchor) and a constant-share residual stack (anchor $(1-b)^L \to 0$) collapse; a restart truncated at depth L (APPNP, and GCNII's initial residual) saturates at its ceiling, and a residual stream with total update $\sigma = 4$ at $e^{\sigma\beta}$. The realized graph-convolution and APPNP/GCNII filters follow the depth and restart curves (Bruce, 2026j). The figure's κ is the collapse ratio κ_{col} of Section 2.3.

4.2 Iterative denoisers and test-time context mixing are the same iteration

The anchored iteration (1.4) of Section 1.4 isolates an update shared by established iterative systems, and Table 4.1 records the scope of each connection. The presence of attention and a residual somewhere in a network is insufficient: the same computation must rebuild its weights and retain a fixed input contribution. For RED, write the published fixed-point update with observation operator B as

$$x_{t+1} = (\sigma^{-2}B^\top B + \lambda I)^{-1}(\sigma^{-2}B^\top A + \lambda f(x_t));$$

setting $B = I$ and $f(x) = \mathcal{P}(x)x$ gives (1.4) with $\alpha = \lambda\sigma^2/(1+\lambda\sigma^2)$, as in Section 1.4. A general inverse problem or arbitrary denoiser does not give this stochastic mixer. The analysis of Chapters 7 and 12 concerns the update itself; it does not assume that $x - f(x)$ is the gradient of a regularizer, an interpretation whose additional conditions are analyzed by Reehorst & Schniter (2019). For p-Laplacian propagation, let D_i be the fixed graph degree, $M_{ij}(F)$ the published reweighted edge, and $c = 2\mu/p$, with μ and p the model's own parameters (Fu et al., 2022, Eqs. 11–17). Put $y_i = F_i/\sqrt{D_i}$ and $a_i = F_i^0/\sqrt{D_i}$. Where $s_i = \sum_j M_{ij}/D_i > 0$, the propagation becomes

$$y_i^+ = \frac{s_i}{s_i + c}\sum_j \frac{M_{ij}}{\sum_k M_{ik}}\, y_j + \frac{c}{s_i + c}\, a_i.$$

Both coefficients are nonnegative and sum to one. For $p \neq 2$, the restart as well as the normalized edges can depend on the state, so the derivative contains additional restart terms; degenerate edge differences also require the model's stated conventions, and smoothness cannot be inferred from the displayed algebra alone. In a DEQ-Transformer the fixed input enters the query, key and value projections at each evaluation of the equilibrium map (Bai et al., 2019, Sec. 4), so the weights can depend directly on A as well as X. For a general update $F(X, A)$, on a differentiable equilibrium

Table 4.1: **Where adaptive mixing and retained input occur together.** Exact algebraic correspondence does not itself establish the contraction hypothesis. Bounds for extensions must include their additional state or input dependence.

System	Evolving weights and retained input	Relation to (1.4)
RED with NLM (Romano et al., 2017; Reehorst & Schniter, 2019)	Current patch similarities; the observed image at every fixed-point step	Exact for identity observation operator and fixed denoiser parameters, with projection inactive.
p-Laplacian GNN (Fu et al., 2022)	Reweighted feature differences; initial features at every propagation step	Convex averaging after degree rescaling, with state- and node-dependent restart for $p \neq 2$.
NeuTRENO (Nguyen et al., 2023)	Recomputed attention; an initial-value fidelity term	Includes $PX + \lambda(A - X)$; $P - \lambda I$ need not be nonnegative.
Nonlinear GRAND++ (Thorpe et al., 2022)	State-dependent diffusivity; fixed, centered input-derived source	Continuous-time diffusion with a selected-node source, not a uniform restart.
DEQ-Transformer (Bai et al., 2019)	Equilibrium-state attention; fixed input injected into attention projections	Includes input-dependent weights, learned transforms, normalization, and feed-forward operations.
APPNP (Gasteiger et al., 2019a)	Fixed graph weights; initial predictions reinjected	Constant-builder special case after degree rescaling; it has no rebuilding feedback.

branch with invertible $I - D_X F$, implicit differentiation gives the input derivative $(I - D_X F)^{-1} D_A F$; the numerator is βI only under the input dependence of (1.4). These distinctions make the common mechanism useful without transferring the certificate of Section 7.1 to an entire architecture unchanged.

Two concrete instances recur in Chapter 12, and this section gives only their settings. The first is test-time context mixing: the anchors are the fixed DINO ViT-S/16 features (Caron et al., 2021) of the CIFAR-10 images (Krizhevsky, 2009) in a context of 32, and fixed affine heads read the mixed features. The experiments mix *images within a context*, isolating the role of adaptive rebuilding; the encoder and classifier stay unchanged (Section 12.4). The second, whose amplification Section 12.3 reports, is a photographic RED/NLM denoiser, which uses Gaussian non-local means in the exact RED denoising specialization. For pixel i, let $\psi_i(X)$ be its 3×3 patch, extended symmetrically at crop boundaries, with Gaussian spatial weights η_k of standard deviation one normalized to sum to one. Within the fixed 7×7 search window $\mathcal{N}_i$, including i and omitting neighbours outside the crop, define

$$\mathcal{P}_{ij}(X) = \frac{\exp\big(-\|\psi_i(X) - \psi_j(X)\|_\eta^2 / h^2\big)}{\sum_{k \in \mathcal{N}_i} \exp\big(-\|\psi_i(X) - \psi_k(X)\|_\eta^2 / h^2\big)}, \qquad j \in \mathcal{N}_i, \tag{4.1}$$

with zero weights outside the window and $\|v\|_\eta^2 = \sum_k \eta_k v_k^2$. The self-similarity before normalization is one. The anchor is a 32×32 crop of the public-domain camera photograph, scaled to $[0, 1]$; no noise is synthesized, and $\alpha = 0.95$, $h = 0.025$. Every update is a convex combination, so all iterates and displayed perturbed inputs stay in the pixel range and clipping is inactive. This is an illustrative dynamics experiment, not a denoising-quality comparison.

Read through this iteration, test-time context mixing has one identity at its centre. Holding the realized weights P_* of an equilibrium fixed gives the frozen smoother $R_* = \beta(I - \alpha P_*)^{-1}$, which is row-stochastic, and the equilibrium is R_* applied to the anchors. For an affine classifier the

output scores are therefore R_* applied to the input scores, and every true-class-versus-competitor margin splits exactly into the sample's own contribution, $(R_*)_{ii}$ times its own input margin, and its companions' contribution, the R_*-weighted sum of their margins for that same class pair (Bruce, 2026h). Section 12.5 states the identity, and Section 12.6 uses it to take apart a focal failure, placement by placement.

4.3 Where gain should sit: placement

Label propagation, restarted propagation, graph diffusion and residual streams all multiply a seed Y (labels or features, one column per channel) by an amplifier $g(P)$, and somewhere they also apply a second linear map M on the token index: a second smoothing kernel, a projection onto a subspace of node functions, the token-mixing part of a readout. The two orders are $Y_{\text{post}} = g(P)\,MY$ (project, then amplify) and $Y_{\text{pre}} = M\,g(P)\,Y$ (amplify, then project), and they differ by the commutator $[M, g(P)]\,Y$ (Section 6.2). This section asks where the gain should sit relative to the second map. The answer does not use the lottery except through one corollary, and it is a worst-case statement: it says how large the difference between the two orders can be, and which order closes which channel. For a single seed y the question is how much of the amplified seed can leave the seed's own direction, because that escaping part is what the two orders treat differently.

Theorem 4.1 (Fragility identity)**.** *Let P have a stationary distribution $\pi > 0$, let $y \neq 0$ be a seed, and write* $\text{Proj}_y x = \langle y,\, x\rangle_\pi\, y/\|y\|_\pi^2$ *and $r = (I - \text{Proj}_y)\,g(P)\,y$ for the escape.*

(i) *On at least two states, the largest seed-orthogonal output is the escape norm:*

$$\max_{z \perp_\pi y,\ \|z\|_\pi = 1} \langle z,\, g(P)y\rangle_\pi = \|r\|_\pi.$$

(ii) *If P is reversible and $w_k = \langle v_k,\, y\rangle_\pi^2/\|y\|_\pi^2$ are the seed's spectral weights, then $\|r\|_\pi = \|y\|_\pi\ \text{sd}_w(g)$, where $\text{sd}_w(g) = \left(\mathbb{E}_w[g^2] - \mathbb{E}_w[g]^2\right)^{1/2}$ is the* gain spread*, the standard deviation of the gain over the seed's spectrum.*

(iii) Amplifying first exposes the escape. *The rank-one map $E = \varepsilon\, u\,(z^\star)^\top D_\pi$, built from the maximizer $z^\star = r/\|r\|_\pi$ (when $r \neq 0$) and any π-unit u, has $\|E\|_\pi = \varepsilon$ and $Ey = 0$; it does damage $\|E\,g(P)\,y\|_\pi = \varepsilon\|r\|_\pi$ under amplify-then-project and none, $g(P)Ey = 0$, under project-then-amplify. No map with $\|E\|_\pi \leq \varepsilon$ and $Ey = 0$ does more damage.*

(iv) Amplifying last leaves the escape in the output. *For a π-unit $n \perp_\pi y$, the projection $M_n = I - n\,n^\top D_\pi$ that removes n gives $\langle n,\, Y_{\text{pre}}\rangle_\pi = 0$ but $\langle n,\, Y_{\text{post}}\rangle_\pi = \langle n,\, g(P)y\rangle_\pi$, which is largest, equal to $\|r\|_\pi$, at $n = z^\star$.*

The proof is in Section A.5; it is Cauchy–Schwarz and an expansion in the eigenbasis. Neither order dominates. Each closes one channel, and both channels have the same worst case, the escape norm $\|r\|_\pi$, which for a reversible operator is the seed's norm times the gain spread. Projecting first means the second map never sees what the amplifier moved off the seed, which protects against a map that reads nuisance the seed does not carry (clause iii); amplifying first lets the second map remove what the amplifier creates, which protects against nuisance manufactured by propagation (clause iv). Which channel matters depends on the map, and here we offer a reading rather than a result: a similarity kernel or learned projection built from a representation that carries the nuisance is exposed to the first channel, so project first; a trusted filter whose job is to remove a direction is exposed to the second, so amplify first. For a seed on a readout mode μ_d and a nuisance mode μ_n with weights w_d and w_n, $\text{sd}_w(g) = \sqrt{w_d w_n}\,|g(\mu_n) - g(\mu_d)|$, and the contamination of the readout

channel under amplify-then-project is multiplied by the *dynamic range* $R = g(\mu_n)/g(\mu_d)$. In a synthetic check on matched perturbations, the damage done by an amplifier with gains above one grows linearly in R (slope 1.002), reaching 495 times the largest damage a filter with gains in $[0, 1]$ can do. The identity has a multiclass form, verified to 1.5×10^{-15} with post-order damage identically zero. For directed P the maximization in (i) is unchanged, but the escape norm contains an output energy, whose derivative at $t = 0$ along the sweep (1.3), $\langle y, [g(S), \mathcal{D}_g]\, y\rangle_\pi$ with $\mathcal{D}_g$ the array of Theorem 3.4(iii) for g, vanishes when P commutes with its time reversal (as in Theorem 3.4(iv)) and generically not otherwise: the escape norm moves at first order in the flow.

The lottery enters through one corollary. An anchor, the $m = 0$ case of the short-path mass that bounds collapse (Theorem 2.8), also bounds how far an amplifier can turn a seed.

Corollary 4.2 (The anchor caps gain ratios and turns)**.** *Let P be lazy and reversible and f a normalized amplifier with anchor $a > 0$. Every gain lies in $[a, 1]$: $a \le f(\mu_k) \le 1$ for all k. Hence $\kappa_{\mathrm{col}} \le 1/a$; the dynamic range between any two modes is at most $1/a$; and for every seed y the angle ϑ between y and $f(P)y$ satisfies $\tan\vartheta = \mathrm{sd}_w(f)/\mathbb{E}_w[f] \le (1-a)/(2\sqrt{a})$, with equality for some seed when 0 is an eigenvalue of P. Without an anchor both are unbounded: for depth, the range $(\mu_n/\mu_d)^L$ grows without bound whenever $\mu_n > \mu_d$, and seeds can be turned arbitrarily close to $90°$.*

Proof. On a lazy spectrum every $\mu_k \in [0, 1]$, and f is nondecreasing there with $f(0) = a$ and $f(1) = 1$; a ratio of two numbers in $[a, 1]$ is at most $1/a$. By Theorem 4.1(ii), $\|r\|_\pi = \|y\|_\pi\, \mathrm{sd}_w(f)$ while the component of $f(P)y$ along y is $\|y\|_\pi \mathbb{E}_w[f]$, so $\tan\vartheta = \mathrm{sd}_w(f)/\mathbb{E}_w[f]$. The gains take values in $[a, 1]$; writing $\bar{f} = \mathbb{E}_w[f]$ for their mean, the Bhatia–Davis inequality (Section A.1, T18) gives $\mathrm{sd}_w(f)^2 \le (1-\bar{f})(\bar{f}-a)$, and $(1-\bar{f})(\bar{f}-a)/\bar{f}^2$ is largest, equal to $(1-a)^2/(4a)$, at $\bar{f} = 2a/(1+a)$, attained by a seed with weight $a/(1+a)$ on the consensus mode and the rest on a mode of eigenvalue 0, whose gain is a. For depth, a seed with weight w on the consensus mode and $1-w$ on a mode of eigenvalue $\mu < 1$, whose gain is μ^L, has $\tan\vartheta \to \sqrt{(1-w)/w}$ as $L \to \infty$. □

A normalized filter never amplifies in absolute terms ($\|f(P)x\|_\pi \le \|x\|_\pi$ for every x, by Lemma 1.2(ii)); what matters for placement is relative amplification, a nuisance mode kept louder than a readout mode, and an anchor caps that at $1/a$. The turn bound is attained on the lazy reversible walk on a ring of four tokens, $\frac{1}{2}I + \frac{1}{4}(Z + Z^\top)$ with Z the cyclic shift, whose spectrum $\{1, \frac{1}{2}, \frac{1}{2}, 0\}$ contains 0, and in a synthetic check a depth-200 filter turns a seed with 1% of its weight on consensus by 84°. (Every normalized filter on a lazy reversible operator has $\mathrm{sd}_w(f) \le \frac{1}{2}$; the anchor's contribution is the bound relative to what the seed retains.)

When several filters compose (heads, stacked operators, a propagation followed by a readout smoother), a natural argument says that if they commute their effects separate, so that any failure to separate is measured by their commutators. What controls separation at a readout mode is instead each filter's misalignment with the frame the readout reads in.

Theorem 4.3 (Composition is controlled by readout misalignment)**.** *Let $\{v_j\}$ be a π-orthonormal readout basis with $v_1 = \mathbf{1}$, and let $G_1, \dots, G_A$ satisfy $\|G_a\|_\pi \le 1$ and $G_a\mathbf{1} = \mathbf{1}$. Split each $G_a = \Lambda_a + H_a$ into its diagonal part in the readout basis and a hollow remainder, and let $h_a = \|H_a\|_\pi$. For $G = G_A \cdots G_1$, the deviation from separability at mode j, $\Delta_j = \langle v_j, Gv_j\rangle_\pi - \prod_a \langle v_j, G_a v_j\rangle_\pi$, satisfies, uniformly in A,*

$$|\Delta_j| \;\le\; \prod_a (1 + h_a) - 1 - \sum_a h_a \;\le\; \tfrac{1}{2}\Big(\sum_a h_a\Big)^2 \exp\Big(\sum_a h_a\Big),$$

with $\Delta_1 = 0$ and, for $A = 2$, exactly $\Delta_j = \langle v_j, H_2 H_1 v_j\rangle_\pi$. There are filters that commute exactly and still have $\Delta_j \neq 0$.

The proof is in Section A.5. The theorem says that "the operators commute, so the analysis separates" is where separation arguments fail; the controlling quantity is each filter's misalignment h_a with the readout's frame. Across 2,000 bound checks on synthetic substrates (random readout bases and flow-perturbed block operators) with two to six filters, the bound was never violated (worst ratio 0.075), and the readout-mode deviation scaled at a mean slope of 1.997 in the misalignment against a mean of 0.862 (median 0.961) for a generic probe.

The fragility identity describes a worst case, and the regime where channel (iii) of Theorem 4.1 does most harm was searched for directly. Its precondition is a nuisance mode kept louder than the readout mode after the consensus mode is deflated, $R > 1$. A first search covered the feature kernels of two spurious-correlation benchmarks (CelebA and Waterbirds) and nine heterophilic benchmark graphs, each in directed and reversibilized form over three seeds: 66 cells. R was defined on 24 of them, which have only eight distinct spectra (three seeds of each kernel, and one graph each for Texas and Wisconsin; each directed variant shares its reversible part, and so the spectrum R reads, with its reversibilized twin); the other 42 were subsamples of the seven larger graphs, which the subsampling left reducible or with a degenerate stationary distribution. None of the 24 entered the regime. The largest deflated range, $R = 0.9965$ on the 183-node Texas graph, occurs at the weakest amplification tested ($\alpha = 0.1$), where every range tends to one, and the margin that does not depend on α, $\mu_n - \mu_d = -0.032$ there, is negative. The feature kernels sat at 0.97–0.99, and the identity itself held to 1.9×10^{-15}. Real operators do enter the regime. In the placement harness of Bruce (2026j), sharp Gaussian kernels on frozen image features (three backbones, two datasets, five seeds) have deflated $R > 1$ in 24 of 30 cells, k-nearest-neighbour graphs with $k = 10$ in 2 of 30 and with $k = 20$ in none, and a spectral gate computed from the seed rejects every sharp-kernel cell. In all 24 sharp-kernel cells both orders lose accuracy against the unsmoothed classifier: at a restart of $\alpha = 0.5$, amplify-then-project loses 0.8 to 4.3 points and project-then-amplify 4.3 to 25.7. In no cell and at no strength did amplify-then-project do worse, so the contamination signature of channel (iii) did not appear. Of the two $k = 10$ cells, one gains about 23 points under both orders and the other is flat. The same harness shows the signature when it is there: a synthetic operator whose nuisance mode ($\mu = 0.99$) is louder than the class modes (0.9) gives project-first gains of 1.2 to 7.4 points at moderate strength ($\alpha = 0.3$–0.7), while at $\alpha = 0.9$ amplify-first loses 1.8 to 6.4 points, decisively in all five seeds. Two caveats limit these searches: the range was reported only after deflation, and weight decay acting as a ridge on the learned features may itself keep R below one. Under project-then-amplify, nuisance carried by the second map cannot reach the output through channel (iii); under amplify-then-project it can, with the worst case of Theorem 4.1. Which kernel constructions enter the regime, and what placement then does in them, is open.

4.4 A strong signal need not be useful: when propagation helps frozen features

Placement decides what amplification can harm. Whether it helps is a separate question.

Proposition 4.4 (Signal-to-noise under amplification)**.** *Let P be reversible with eigenpairs (μ_k, v_k), and let the seed be $x + u$ with a fixed signal x of spectral weights $w_k = \langle v_k, x\rangle_\pi^2 / \|x\|_\pi^2$ and a random noise u that is isotropic in the walk's metric, $\mathbb{E}\langle z, u\rangle_\pi^2 = \omega^2 \|z\|_\pi^2$ for every z. Then $\mathbb{E}\|f(P)u\|_\pi^2 = \omega^2 \sum_k f(\mu_k)^2$, and amplification multiplies the signal-to-noise ratio by*

$$\frac{\mathbb{E}_w[f^2]}{\frac{1}{N}\sum_k f(\mu_k)^2} :$$

it helps exactly when the signal sits where the squared gain exceeds its average over the spectrum.

Proof. $\mathbb{E}\|f(P)u\|_\pi^2 = \sum_k f(\mu_k)^2\,\mathbb{E}\langle v_k, u\rangle_\pi^2 = \omega^2 \sum_k f(\mu_k)^2$, while $\|f(P)x\|_\pi^2 = \|x\|_\pi^2 \mathbb{E}_w[f^2]$ and $\mathbb{E}\|u\|_\pi^2 = N\omega^2$. □

The proposition is a precondition: amplification cannot help a seed whose signal does not sit where the gain is high. It assumes noise that is isotropic in the walk's metric. Noise that comes from sampling which nodes are labelled is not isotropic when π is not uniform; that case, and the full set of measurements of propagation on frozen features that test the precondition, are left for later work. Two of the measurements bound what this section claims. On median-bandwidth Gaussian graphs of frozen CIFAR features the class modes are barely separated from the rest of the spectrum (gap 4.7×10^{-4}, measured by a gauge whose record is not included), and a restart at $\alpha = 0.9$ lowered accuracy under both orders alike in all 60 cells (three backbones, two datasets, ten seeds), by about 40 points at the median.

The second shows that, because the seed, not the operator, carries the signal, one operator can help one seed and hurt another. On fifteen k-nearest-neighbour graphs of frozen CIFAR-100 features, from three backbones with five seeds each, propagating a classifier's predictions after the classifier (project, then amplify, as in correct-and-smooth, Huang et al., 2021), with the labelled set as the only moving part, the smallest change among the post-placement runs using restarts at $\alpha = 0.3, 0.5, 0.7$ and one-step smoothing alone or after the $\alpha = 0.5$ restart was negative on every graph at 40 labels per class (80% of the nodes labelled), from -0.1 to -3.9 points, and positive on every graph at 4 per class (8%), from $+7.6$ to $+10.9$: a paired within-graph swing of $+10.9$ points (95% paired-bootstrap interval $[10.2, 11.6]$ across the fifteen graphs) at 8% against 80% (the registered primary, pooled over both low densities, is $+8.8$ points, with the corresponding interval $[8.2, 9.4]$). The effect is monotone in label density (4, 8 and 40 per class) on all fifteen graphs. The graph is the same at every density and the labelled sets are nested, so the harm turns into a gain on identical graphs. The missing premise is headroom, the signal a propagation step can still import; this measurement does not separate headroom from the coverage of nodes with no labelled neighbour. Chapter 0 counts the flip among the results of the collection (Section 0.4.6).

What, then, do the spectral gauges certify? The best of them, the constant-mode-deflated advantage, a gauge of signal computed from the seed, predicts gain out of sample more strongly than the task gap at the class-count boundary in the spectrum. On a battery of 61 graphs, seven of them real benchmark graphs and the rest synthetic, gain is APPNP's test accuracy minus a raw-feature linear probe's, averaged over three training seeds per graph. Monotone predictors fitted on 31 of the graphs and tested on the other 30 have Spearman correlations with the measured gain of 0.72 for the advantage and 0.21 for the gap; this is one split, and an exploratory comparison. But the advantage rises with label density: on the fifteen graphs above its mean is 1.15, 1.37 and 3.09 at 4, 8 and 40 labels per class, and the gauge admitted all fifteen graphs at 40 labels per class and none of the thirty at 4 or 8. It admits exactly the high-density cells, whose reported minimum post-placement effect is negative, and rejects the low-density cells, whose reported minimum is positive. A certificate of signal is not a certificate of utility: utility is signal times headroom (Section 0.4.6).

Where does the harm live? On k-nearest-neighbour graphs of frozen image features from three backbones, with 80% of the nodes labelled, propagating after the classifier was harmful, by -1.0 to -4.6 points, in 16 of the 26 cells that the spectral gauge admitted, while every positive control showed the pattern it was built to show. Fifteen of these graphs are the ones of the label-density sweep: at 40 labels per class it reproduces their results, and at 4 and 8 the harm is gone, so on these graphs the harm lives where labels are dense, the case the gauge admits. The other harm, contamination of the readout through channel (iii) of Theorem 4.1, was not found: in the initial benchmark search, no case with a defined range entered its regime, and where sharp kernels did enter it, its signature did not appear (Section 4.3). The flip is measured only on neighbour graphs

built from frozen image features; on real graphs it has not yet been measured.

For practice, the measurements of this section give one rule, the one Chapter 0 states: utility is signal times headroom (Section 0.4.6). A gauge of signal alone does not decide whether to propagate; what propagation can still import has to be weighed with it. A headroom-aware rule that predicts utility across settings is left for later work.

Pending: The headroom flip on a second dataset and backbone: propagation gain against label density on neighbour graphs of CIFAR-10 and ImageNet-100 features, with the deflated advantage and the cells the gauge admits.

4.5 What the graph brings: consistency amplifies class structure

Pending: How graph-consistency regularization turns a response to perturbation into class structure: the neighbour graph is the operative factor and the perturbation secondary, and the gain depends on cluster structure, headroom and balance.

4.6 What holds in trained networks

Shown in an idealized setting. The same laws in graphs, denoisers and test-time mixing (Sections 4.1 and 4.2): the collapse law is stated for powers of one fixed operator, which linear propagation over a fixed graph satisfies exactly; the certificate is stated for one builder $P(\cdot)$ applied repeatedly, as in iterative denoisers, deep equilibrium models and test-time mixing. A transformer's forward pass applies a different $P(X)$ at each layer.

Measured in trained networks. A strong signal need not be useful (Section 4.4): measured with frozen features and label propagation on neighbour graphs.

The theorems that Chapters 1–4 take from Bruce (2026i) are proved; the measurements test their reach. Machine-precision checks (ID) certify that identities are implemented correctly and hold on real operators, and the synthetic checks (SYN) test conditions and constants; neither is evidence about trained networks. Checks on real operators or spectra along an artificial perturbation are marked SYN-R, and measurements of trained models and real data REAL. Table 4.2 lists what was measured on trained models and real data, with the caveat that limits each row; Appendix B lists the quoted numbers and their sources.

Table 4.2: What was measured on trained models and real data, and what limits each measurement.

Claim	**Result**	**Tier**	**Main caveat**
Sink weight rarely vanishes	median δ 0.416 (7 checkpoints), 0.080 (GPT-2); 0.67%, 2.2% below 10^{-6}	REAL	two protocols that disagree on the median; structural sink, not functional
Single layers collapse fast	median half-collapse 1–2 applications	SYN-R	iterating one captured layer is not a trajectory
Stack clock	GPT-2 median composite distance to the sink 0.76 at $b = 0.2$, 0.090 at $b = \frac{1}{2}$	SYN-R	b assumed; value, output and MLP maps omitted
Two readouts at native flow	median collapse-type change 0.2–1.5%; median eigenvector turn about 16–17°	REAL	lazy, head-averaged; 16 texts, 8 image crops; two small encoders
Words of layers	median slopes 2.000 (palindromes) and 1.000 (non-palindromes)	SYN-R	layers balanced to a shared π
Signed filters for relevant modes at low or negative eigenvalues	Mean gains over nonnegative filters: +24.3, +28.2 pp on Texas, Wisconsin; +3.9, +1.9 on Squirrel, Actor	REAL	small heterophilic graphs; matched budgets; 10 seeds per model
Without a task band there is no gain	both orders lose in all 60 cells at $\alpha = 0.9$ (median about 40 pp)	REAL	one graph construction (median-bandwidth Gaussian)
Label density decides	smallest post-placement change: harm at 80% labels, gain at 8% on all 15 graphs; paired mean swing +10.9 pp at 8% vs. 80% (registered primary, both low densities: +8.8 pp)	REAL	one dataset family; headroom and coverage not separated here
Placement null, first search	$R \leq 0.9965$ on the 24 defined cells	REAL	eight distinct spectra; 42 of 66 cells undefined; ridge confound
Harm regime on real features	deflated $R > 1$ in 24/30 sharp-kernel cells; both orders lose; no channel-(iii) signature	REAL	one kernel family on frozen image features
No collapse without mixing	an MLP does not collapse across samples: depth 4 only (8–64 did not converge)	REAL	negative control only

Part II

What attention needs

Chapter 5

Short paths: the anchor

Part I defined the object and its laws. This part asks what the object needs. An averaging layer that is rebuilt from its input and iterated faces six failure modes, and Chapter 0 lists them with the protection that answers each (Section 0.3.2): an anchor, a free content path, gain control, a way to abstain, a routing budget that grows with length, and an arrow. A seventh condition, a writer, stands outside that inventory, because moving information around is not enough. Chapters 5–11 take the seven in turn. Each states its condition as a theorem within a declared class, gives the signature measured when the protection is removed, asks whether capacity elsewhere can compensate for its absence, and says how the standard pre-normalized block pays for it. These are conditions on a class of models, not mechanisms that every successful model must implement: they say what goes wrong, and at what price, when a model in the class lacks the property. Table 0.2 in Chapter 0 summarizes the part.

The first condition is the one Chapter 2 proved. Its retention theorem and its collapse-ratio theorem are one statement in two forms (Sections 2.2 and 2.3): a mode with $0 < \mu < 1$ survives arbitrary depth if and only if a fixed share of paths stays within a bounded number of steps. The anchor is what supplies that share. This chapter states the condition and the remedies that already keep it under other names (Section 5.1), gives its removal signatures (Sections 5.2 and 5.3) and its form for routers that copy rather than average (Section 5.4), shows why, under strict mixing, capacity elsewhere of bounded gain cannot buy it back (Section 5.5), and marks where its class ends (Section 5.6).

5.1 The anchor under its many names

Let $f(P) = \mathbb{E}[P^K]$ be a normalized amplifier of a lazy reversible operator with content eigenvalue $\mu_C = 1-\beta \in (0,1)$. Collapse stays bounded as the filter is strengthened if and only if the path-length law keeps uniform short-path mass, $\Pr[K \le m] \ge \rho$ for some fixed m and $\rho > 0$ (Theorem 2.8). The same criterion decides the retention of any interior positive mode of a fixed operator (Theorem 2.1). An identity or restart path supplies short-path mass, but it is sufficient, not necessary: the law $K \equiv 1$ retains with no identity atom. A residual stack with a fixed mixing share p still collapses, only more slowly. Its gain $(1 - p(1-\mu_C))^L$ falls tenfold at about $L \approx 2.3/(p(1-\mu_C))$ layers when $p(1-\mu_C)$ is small; plain powers take about $2.3/\beta$ layers when β is small. When the per-layer gains stay bounded below, the gain is $\exp(-\Theta(S_L))$ in the mixing budget S_L, the expected number of layers that mix, so a floor needs a bounded budget (Theorem 2.11).

Read through Theorem 2.8, remedies developed independently in several literatures keep short-path mass in different guises (Table 5.1). The theorem does not say that residual connections are

Table 5.1: Independently developed remedies read through Theorem 2.8. Each keeps short-path mass through a share of the lottery at $K = 0$ (for the residual stream, when its shares stay away from one); GCNII's identity mapping acts on the feature side and is not the anchor, and the harmonic clamping of Zhu et al. (2003) is an absorbing boundary rather than a restart.

System	**The anchor's guise**	**Lottery**	**Anchor** a
Residual stream with bounded total update $\sum_\ell b_\ell$	the $\mathcal{S} = \emptyset$ term: the input itself	independent Bernoulli sum	$\prod_\ell (1 - b_\ell)$
APPNP / personalized PageRank (Gasteiger et al., 2019a)	teleport to the input	geometric (truncated)	teleport probability
GCNII's linear propagation, with identity feature maps and a fixed initial-residual weight (Chen et al., 2020a)	initial residual to $H^{(0)}$	geometric (truncated)	initial-residual weight
Label spreading (Zhou et al., 2003)	the seed labels as a restart prior	geometric	$1 - \alpha$

only an anchor, since they also carry gradients (He et al., 2016); it says that within the filters depth produces, short-path mass is the only way to keep collapse bounded, and that a residual stream supplies it when its total update is bounded. In the scalar-share idealization, the constant shares of standard residual stacks (He et al., 2016; Vaswani et al., 2017) slow collapse but do not bound it. Chapter 0 reads the usual description of the residual stream as a bus, which each layer reads and to which it adds an edit (Section 0.2.2), through this condition: the residual is the anchor: of the deficits tested so far, the one that nothing else in the block made up for, and under strict mixing no compensator of bounded gain can (Table 0.1; Section 5.5).

5.2 Removal signatures

Removing the identity path sends a minimal attention block (Section 16.6) to exact chance at depth 12 and beyond, measured on trained models (Bruce, 2026d). In a 22M-parameter block, removing the attention sublayer's identity path gives exact chance, with a constant-ranking signature in the records. Neither outcome isolates the collapse mechanism. The collapse horizon of plain graph-convolution operators is met on schedule, on real operators along an artificial perturbation; it is an operator quantity, and trained GCNs collapse far earlier, at depth 8–16, while the shortest operator horizon, on a Cora subgraph, is near 87.

Post-LN, whose mixing budget grows linearly with depth (Section 2.6), shows a trainability cliff at depth 18 and beyond on CIFAR-10, bimodal across seeds, measured in small trained vision transformers (Bruce, 2026d).

Looped models lose depth extrapolation without input re-injection: a weight-tied block trained at 4 loops keeps 0.023–0.032 accuracy at 96 loops without it, against 0.904–0.959 with it (three seeds, Table 5.2), measured on trained models (Bruce, 2026d).

The comparison at 22 million parameters, as recorded (Section 16.7), includes deletions from a minimal block of depth 12, three seeds each, one of which removes the attention sublayer's identity path (Section 16.6). Without it, top-1 accuracy is 0.5% on Tiny-ImageNet's 200 classes and 1% on ImageNet-100, chance on both, paired drops of 45.27 and 54.87 points over the deletion seeds; the intact block averages 45.92% and 55.96% over its five seeds. These are tabulations of recorded results, not a reproduction. In the small minimal block of the first signature, on CIFAR-100, the identity-free arm still learns at depth 4 (10.02–12.66% top-1, against 31.03–33.11% intact) and sits

Table 5.2: **A looped block keeps its accuracy at depth only when its input is re-injected.** One weight-tied Pre-LN causal attention-plus-MLP block, applied K times, on multi-query associative recall (MQAR), trained at K_{train} loops with no re-injection (none), with the input re-injected at every loop (reinj), or with an exponential moving average of the latent that starts at the input (ema); a_K is the accuracy at K loops, per seed. 2,048 answers per value, chance 1/64. The conservative sequence-level 95% interval half-width is at most 0.063. $K_{1/2}$ is the loop count at which accuracy falls to half its value at K_{train}; horizons are rounded, with their bracketing grid points, and > 96 is right-censored. The untrained cell is italicized and excluded from horizon comparisons. At $K_{\text{train}} = 64$, none of the six runs without the moving average trained, and three of the six with it.

Arm	K_{train}	Seed	$a_{K_{\text{train}}}$	a_{64}	a_{96}	$K_{1/2}$ [bracket]
none	4	0	1.000	0.049	0.032	21 [16, 32]
none	4	1	1.000	0.046	0.031	21 [16, 32]
none	4	2	1.000	0.033	0.023	20 [16, 32]
none	16	0	1.000	0.621	0.228	72 [64, 96]
none	16	1	1.000	0.444	0.183	60 [32, 64]
none	16	2	0.990	0.522	0.225	66 [64, 96]
reinj	4	0	1.000	0.933	0.904	> 96
reinj	4	1	1.000	0.972	0.959	> 96
reinj	4	2	1.000	0.954	0.939	> 96
reinj	*16*	*0*	*0.207*	*0.189*	*0.184*	*did not train*
reinj	16	1	1.000	0.999	0.997	> 96
reinj	16	2	0.999	0.997	0.994	> 96
ema	4	0	1.000	0.156	0.077	35 [32, 64]
ema	4	1	1.000	0.270	0.134	44 [32, 64]
ema	4	2	1.000	0.171	0.087	36 [32, 64]
ema	16	0	1.000	0.986	0.884	> 96
ema	16	1	1.000	0.983	0.843	> 96
ema	16	2	1.000	0.989	0.835	> 96

at exactly 1.0%, chance, at depths 12 and 24 in all five seeds. Whether any capacity elsewhere can stand in for the removed path is the sweep that Section 5.5 reports as pending.

5.3 Pre-LN criticality and the Post-LN ledger

Normalization placement is a choice of mixing budget (Section 2.6). In the mean-field energy model at initialization, Pre-LN sits on the critical line between a floor and exponential loss, with a budget that grows like $s \log L$, where s is the attention share (Proposition 2.21). Post-LN resets the stream after every layer, so its budget grows linearly with depth, and the trainability cliff observed from depth 18 in small vision transformers on CIFAR-10, bimodal across seeds, is the removal signature of Section 5.2.

In trained stacks the anchor can be read off layer by layer as a ledger. Each layer's anchor coefficient $\hat{a}_\ell$ is measured from its input and output, and the ledger $\sum_\ell (1 - \hat{a}_\ell)$ mirrors the total anchor defect of the products clause of Theorem 2.8. In graph transformers trained on ten graph datasets at depths 2 to 16 (Section 15.4), a pre-normalized stack's mean last-layer anchor coefficient is 0.998, against 0.934 for Post-LN, over 165 cells per placement using five split/seed indices with incomplete depth coverage (Section 0.3.3) , and across depth the Post-LN ledger grows by about 0.058 per layer while the fitted Pre-LN slope is slightly negative. The measured anchor is an input–output regression coefficient; it is not a path probability. Whether the ledger predicts Post-LN's

test-score deficit is not settled. The score is ROC-AUC on Minesweeper and Tolokers and accuracy elsewhere. The pooled correlation is 0.546 over the 165 matched Post-LN pairs recorded so far, of the 200 a complete grid would give, against a threshold of 0.70 set in advance; Post-LN's mean test score drops by at least three percentage points from depth 4 to 16 on only one of the seven fully scored datasets ; and the study's joint registered outcome is inconclusive. The positive pooled association depends on the cohort: a post-hoc check finds a negative correlation on the fully covered accuracy-only subset. Chapter 0 reads the usual answer "it trains: conditioning is half the design" (Section 0.2.2) through two conditions: "it trains" becomes the anchor, critical at initialization in Pre-LN, and gain control (Table 0.1; Chapter 7).

5.4 The tether: consensus and coalescence

A machine that rebuilds every token from what it routes to it can lose the tokens themselves. Soft routing averages them toward one another. Hard routing copies one token onto another until whole groups share a state. A tether, a path back to each token's own input, prevents both, and what it costs depends on how concentrated the routing is.

Proposition 5.1 (The tether criterion)**.** *Let P be a fixed row-stochastic matrix and $0 \le \alpha < 1$. The tethered iteration $X_{t+1} = (1-\alpha)X_0 + \alpha P X_t$ has the unique fixed point RX_0, with $R = (1-\alpha)(I-\alpha P)^{-1}$, and it attracts every starting state.*

(a) If the rows of X_0 are unit vectors with Gram matrix G, nearest-cosine decoding returns token i's own identity, with no tied competitor, exactly when $(RG)_{ii} > (RG)_{ij}$ for every $j \neq i$. For orthonormal rows this is $R_{ii} > \max_{j \neq i} R_{ij}$.

(b) With orthonormal inputs and hard routing, $P_{i,f(i)} = 1$, a token on a cycle of length m keeps its identity for every $\alpha < 1$. A token $d \ge 1$ steps before a cycle of length m keeps it exactly when $\alpha^d + \alpha^m < 1$: $\alpha < (\sqrt{5}-1)/2 \approx 0.618$ for $(d,m) = (1,2)$ and $\alpha < 1/\sqrt{2} \approx 0.707$ for $(d,m) = (2,2)$.

(c) With orthonormal inputs, on a two-cycle, after t rounds from the token's own input, the token keeps its identity at every even t and at odd t exactly when $1 - \alpha > 2\alpha^{t+1}$.

The proof is in Section A.6.

R is the restart's lottery operator, the geometric lottery of Table 1.1 with continuation probability α: a token's output is its own input with weight $1-\alpha$, plus everything that returns to it along paths of routing. With P fixed, the iteration is the model of social influence of Friedkin & Johnsen (1990), in which every actor stays anchored to its initial opinion; its anchor-free limit $\alpha \to 1$ is the consensus model of DeGroot (1974). Its own weight is at least $1-\alpha$, and return paths raise it. Soft routing spreads its mass over many tokens, so the own weight wins easily. Hard routing concentrates it, so the criterion bites. The routing graph of a hard router is a functional graph: every token points to exactly one token, and the paths end in cycles. For a symmetric, strict, self-excluding nearest-neighbour score the cycles have length two, because a longer cycle would need its edge scores to rise all the way around it. Ties or asymmetric scores allow longer cycles, and self-routing allows cycles of length one. Tokens on a cycle keep their identity at every α. Tokens that feed a two-cycle need a stronger tether: $\alpha < 0.618$ one step away and $\alpha < 0.707$ two steps away. These thresholds compare entries of R, so they decide identity only under the decoding premise of part (a), which is why parts (b) and (c) assume orthonormal inputs. Two identical inputs can never be told apart, and near-orthogonal inputs need a margin; Section A.6 gives a sufficient one in terms of the coherence $\max_{i \neq j} |G_{ij}|$. Without a tether the two kinds of router fail differently. Soft routers reach

consensus: every token converges to one state. Hard routers coalesce: tokens collapse onto their cycles while the population as a whole stays spread out. A collapse measure based on spread sees the first and misses the second. For machines that are not pure smoothers, the measure that sees both is identity retention, the fraction of tokens that still decode to themselves. The collapse ratio κ_{col} of Section 2.3 is defined for soft averaging operators; a hard router lies outside that scope rather than contradicting it.

Constructed-key records. A constructed tether test routes 256 product-key tokens carrying 64-dimensional random cargo for 64 rounds, with three seeds, through three routers: softmax, a walker that routes by sampling, and a hard tree that sends each token to a single other token (Figures 5.1 and 5.2) (Bruce, 2026e).

- *No tether.* Softmax identity is at most 0.012 by round 8, and every seed ends in a single state. The walker's spread stalls at its sampling-noise floor, about 0.12. The hard tree's identity equals its two-cycle fraction at even rounds, 0.64–0.77, and is zero at odd rounds, while its spread stays at least 0.997 with 164–198 distinct states. It has coalesced, and its spread does not show it.
- *With a tether.* At $\alpha = 0.9$, softmax keeps every token's identity from the second round on, and the walker keeps 0.83–0.94. The tree keeps exactly its two-cycle tokens: every token that feeds a cycle in this graph fails the condition of Proposition 5.1(b) at $\alpha = 0.9$. At $\alpha = 0.7$ it keeps 0.023–0.047 more than that. The criterion predicts these to be tokens two or more steps from their cycle, which it admits at this α, two steps away just below its threshold 0.707. The cargo is random and 64-dimensional, not orthonormal, so cross-talk decides some of them, and the criterion ignores it. At $\alpha = 0.5$ every tree token keeps its identity.

The accepted recovery now supplies class-resolved measurements (Figure 5.1). The saved class and odd-round checks pass in all three seeds. At $\alpha = 0.7$, 26 of the 46 deeper tokens retain identity: this is the recorded decode, not the larger count predicted by the orthogonal criterion. The saved Gaussian expectation describes finite-cargo cross-talk, not an interval estimate. The recovery's saved overall assessment and execution gate remain failed; its registered exception permits publication as a rebaseline, not historical reproduction. The effective-rank differences still violate zero tolerance. Agreement with the frozen recovery does not erase their historical mismatch, and undefined historical endpoint references are admitted only for the enumerated, replay-verified roundoff defects; they are not called reproduced (Bruce, 2026e, Appendix C.1).

5.5 Why capacity cannot buy back the anchor

Necessity theorems are exact on their own constructions, and trained systems evade them by composition wherever another resource has room, so the neighbourhood of the standard design is flat (Section 15.1). The anchor is the exception. Under strict mixing, recovering lost content needs gain that grows with the collapse ratio, so no compensator of bounded gain buys back short-path mass, although at finite depth large gain can. Of the deficits tested so far, it is the one that nothing else in the block made up for; a test that adds capacity is pending. Chapter 0 states the bound behind it (Section 0.3.4): if the layers act on a common discriminative invariant subspace with norms $q_\ell \le 1$, any downstream recovery map with Lipschitz constant G that must separate two distinct inputs in that subspace by $c > 0$ times their distance needs $G \ge c/\prod_\ell q_\ell$ when the product is positive; large gain at finite depth, an operator close to the identity or an external input escapes it.

Strict mixing has a sharp form for attention whose kernels change from layer to layer, sink-like kernels included. The common Euclidean centered norm need not contract a sink-like stochastic

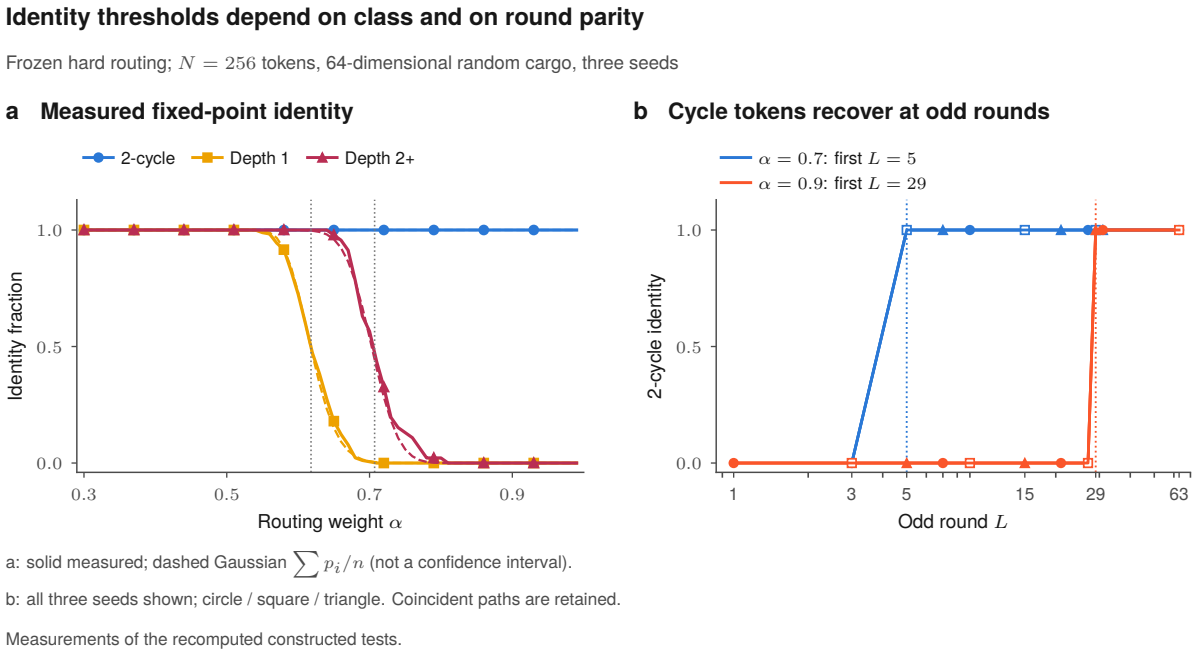


Figure 5.1: **Recovered identity measurements resolve token classes and the odd-round transient.** Frozen hard routing uses $N = 256$ product-key tokens, 64-dimensional random unit-vector cargo and three seeds. (a) Solid curves are saved pooled fixed-point identity fractions for 544 cycle, 178 depth-one and 46 deeper tokens across the seeds. Dashed curves show the saved Gaussian expectation $\sum_i p_i/n$, not uncertainty; deterministic ties with undefined expectation are omitted from that curve. Dotted references mark $(\sqrt{5}-1)/2$ and $1/\sqrt{2}$. (b) All three seeds at $\alpha = 0.7, 0.9$, with circle/square/triangle markers and no offsets; coincident paths are retained. The first identity-preserving odd rounds are $L = 5, 29$, as predicted by $1 - \alpha > 2\alpha^{L+1}$. The saved class and odd-round checks pass. No confidence intervals are implied. The saved execution failure remains and historical records are not reproduced; the registered rebaseline admits these constructed measurements (Bruce, 2026e, Appendix C.1).

matrix. Row diameter supplies a sharp substitute, valid for vector features, shared nonexpansive writers, and realized adaptive routing.

Definition 5.2 (Setting)**.** A state X has rows $x_1, \dots, x_n$, one per token, and a mixing step multiplies it on the left by a row-stochastic matrix P, so $P\mathbf{1} = \mathbf{1}$. The state is at *consensus* when its rows are all equal. For Euclidean features the *centered Euclidean norm* of P is $\|HPH\|_2$, where $H = I - \mathbf{1}\mathbf{1}^\top/n$. We call a downstream map a *compensator* when it reconstructs the desired task from the state actually retained. *Re-supply* additionally introduces information through an input, memory or bypass not contained in that state.

Theorem 5.3 (Sharp vector-valued Dobrushin coefficient)**.** *Let P be any finite row-stochastic matrix and let X have rows in a normed vector space. Define*

$$d(X) = \max_{i,j} \|x_i - x_j\|, \qquad \tau(P) = 1 - \min_{i,j} \sum_k \min(P_{ik}, P_{jk}) = \frac{1}{2} \max_{i,j} \sum_k |P_{ik} - P_{jk}|.$$

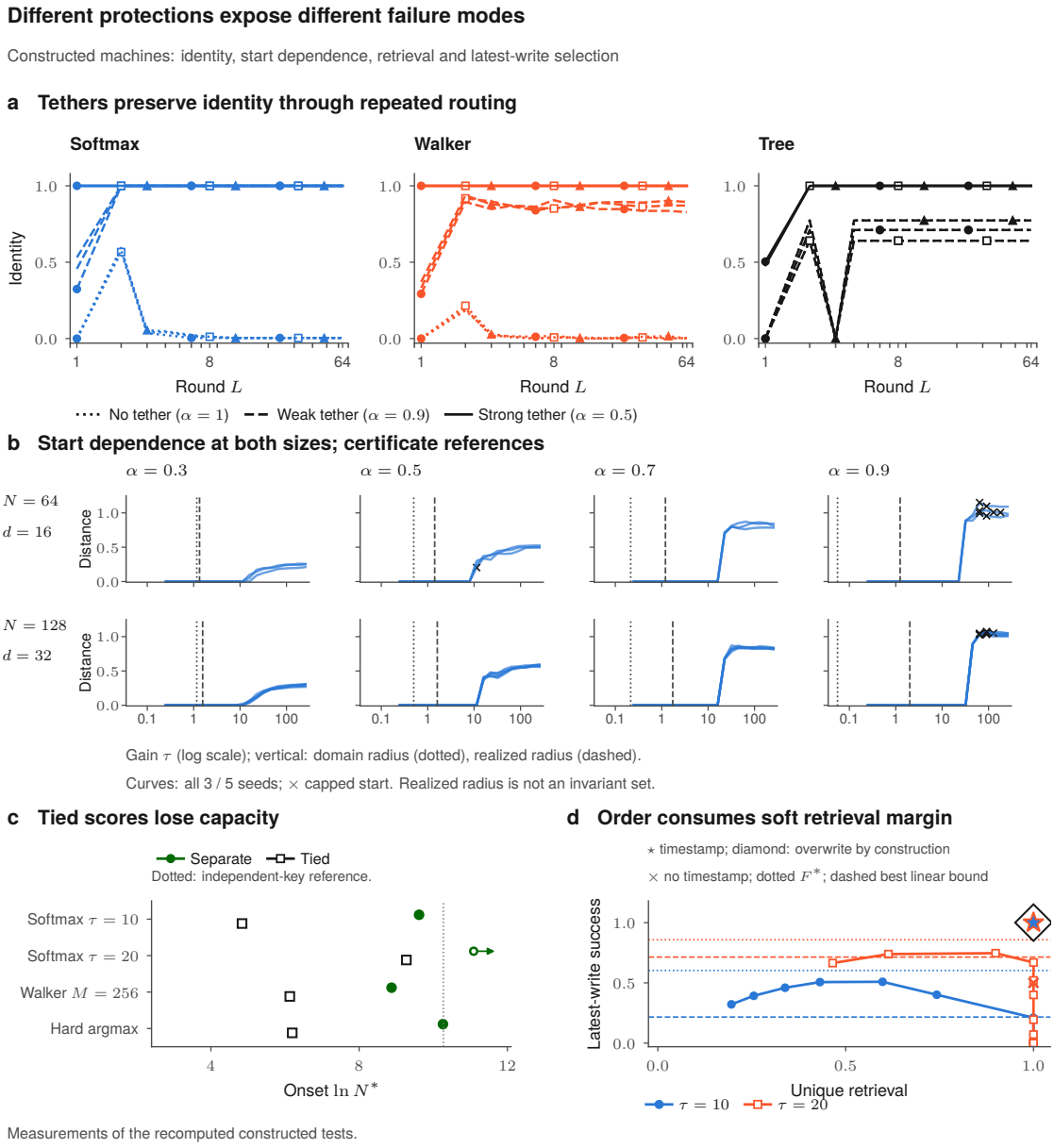


Figure 5.2: **Removing different protections exposes different failure modes in constructed machines.** (a) Identity over routing rounds for $N = 256$, 64-dimensional random cargo, $\tau = 10$ and 64 walkers: all three seeds at $\alpha = 1, 0.9, 0.5$; markers distinguish seeds. (b) Relative endpoint distance $\|Y - X\|_F / \|X\|_F$ for softmax at $N = 64, d = 16$ (three seeds, 1,500 updates) and $N = 128, d = 32$ (five seeds, 5,000 updates), at $\alpha = 0.3, 0.5, 0.7, 0.9$; the larger test retains its 1,500-step states. Crosses mark either capped start; endpoint separation is not a fixed-point certificate. Vertical references are saved median tight certificates at the domain and realized radii; the latter is not an invariant set. (c) Saved tied/untied retrieval onsets; the arrow is right censoring, not an interval. The independent-key reference, 10.274 nats, is the argmax's evidence wall at $c = 0.35$, the cosine of the walker and hard-argmax rows. (d) All 12 recency biases at each gain, timestamp/no-timestamp walkers, overwrite by construction, and saved mean-field references conditional on retaining every position, not empirical maxima. No seed intervals exist for the pooled retrieval curves. The source paper gives protocols and both certificate units (Bruce, 2026e, Table 5 and Appendix C.1). Saved execution failure and historical non-reproduction remain; publication is an admitted rebaseline.

Then $d(PX) \le \tau(P)d(X)$. The coefficient is the smallest possible uniform coefficient even for scalar features. Also $\tau(PQ) \le \tau(P)\tau(Q)$ for stochastic P, Q.

The proof is in Section A.6. The coefficient is Dobrushin's, $\tau(P) = 1 - \lambda(P)$ with λ the uniform leak of Definition 2.17.

Corollary 5.4 (Shared feature maps and residual escape)**.** *Suppose a common linear feature map W_ℓ has operator norm at most w_ℓ, and a common rowwise map ϕ_ℓ is L_ℓ-Lipschitz. For $F_\ell(X) = \phi_\ell(P_\ell(X)XW_\ell)$ one has the pathwise bound $d(F_\ell(X)) \le q_\ell(X)d(X)$ with $q_\ell(X) = L_\ell w_\ell \tau(P_\ell(X))$. For $0 \le p_\ell \le 1$ and $X_{\ell+1} = (1-p_\ell)X_\ell + p_\ell F_\ell(X_\ell)$,*

$$d(X_L) \le \prod_{\ell<L} \left[1 - p_\ell + p_\ell q_\ell(X_\ell)\right] d(X_0).$$

If every realized $q_\ell \le \bar{q} < 1$, this is at most $e^{-(1-\bar{q})\sum_{\ell<L} p_\ell} d(X_0)$. Thus an unbounded total mixing share forces consensus in diameter under these premises. The constant-factor bound is sharp.

The proof is in Section A.6. In particular $w_\ell, L_\ell \le 1$ give exactly the version for shared nonexpansive writers. The bound is on within-state row differences even if P depends on X; it is not a Lipschitz bound between two different inputs, which would also involve the change in P. Different token-specific writers or positional inputs need a separate term and do not satisfy the shared-map premise automatically. The total mixing share $\sum_\ell p_\ell$ plays the part of the mixing budget of Section 2.4.

Worked example 5.5 (Diameter contracts although the centered Euclidean norm exceeds one)**.** Take four rows, the first three of P equal to $e_1^\top$ and the last equal to $(1-q)e_1^\top + qe_2^\top$, with $q = 0.9$. Then $\tau(P) = q < 1$. For $H = I - \mathbf{1}\mathbf{1}^\top/4$, $HPH = q(He_4)(e_2 - e_1)^\top$, so $\|HPH\|_2 = q\sqrt{3/2} > 1$. This example demonstrates why the diameter statement is useful for a sink-like kernel.

Proposition 5.6 (Re-supply and a downstream gain requirement)**.** *If an additional state injection U_ℓ is added at step ℓ, set $u_\ell = d(U_\ell)$ and $a_\ell = 1 - p_\ell + p_\ell q_\ell$. Then*

$$d(X_L) \le \Big(\prod_{j<L} a_j\Big) d(X_0) + \sum_{\ell<L} \Big(\prod_{j=\ell+1}^{L-1} a_j\Big) u_\ell.$$

Without re-supply, write $r_L = \prod_{j<L} a_j$. A common downstream G-Lipschitz row decoder that separates some output pair by $\Delta > 0$ requires $G \ge \Delta/[r_L d(X_0)]$ when the denominator is positive. If that denominator is zero, such separation is impossible for any common row decoder.

The proof is in Section A.6. Consensus does not imply that a pooled global task has lost every piece of information. This gain requirement concerns token distinction under the specified shared readout. Nor does bounded total mixing guarantee retention: a single erasing branch can already destroy it.

Remark 5.7 (Continuity and notation). No bounded-gain reconstruction follows from injectivity alone: $A_\epsilon = \mathrm{diag}(1, \epsilon)$ preserves information for $\epsilon > 0$ but recovering the second coordinate requires gain at least $1/\epsilon$. For each fixed $\epsilon > 0$ this linear map has a continuous linear inverse; the example proves only that no *uniformly bounded-gain* reconstruction follows from injectivity as $\epsilon \downarrow 0$. Its second-coordinate gain is $1/\epsilon$. More generally a full-column-rank finite-dimensional matrix has a bounded left inverse on its image. A product of contraction factors such as $\prod_\ell q_\ell$ is a contraction multiplier, the reciprocal of a collapse ratio in the convention of Section 2.3. We call the contraction multiplier r_L, not κ, to avoid that ambiguity.

Pending: Whether any capacity can replace a removed anchor: with the identity path removed, MLP width, heads and depth are swept in trained networks against arms that re-supply an anchor another way.

5.6 The class boundary: depth without a residual

The anchor theorem is a statement about averaging: normalized nonnegative path laws of one operator (Theorem 2.8). A mixer whose token transport is an orthogonal map rather than a nonnegative average lies outside that class, and there depth without a residual is possible. On a synthetic task in which each of 32 tokens must copy its partner's content, orthogonal pure routing reaches deeper without a residual than attention does, but outcomes depend on seeds and on the training recipe (Figure 5.3). Without a residual, attention is at chance from depth 4 on in all three studies. With the value and output weights constrained to be orthogonal, pure routing keeps 0.977–0.982 accuracy at depth 4 and about 0.88 in two of three seeds at depth 8, while orthogonal weights do not rescue a mixed block that averages a grouping and a conservative routing. With a tuned recipe and per-block LayerNorm off, the deepest tested depths at which single seeds of pure routing reach 0.5 are 24, 32 and 24, but only 8, 32 and 8 when every shallower tested depth must also pass. Budgets and initialization differ across the studies, so these are descriptive comparisons, not isolated causal effects. Chapter 0 counts invertible transport among the currencies in which content survives depth (Section 0.6.3), and Chapter 11 takes up what orthogonal transport costs.

5.7 What holds in trained networks

The anchor's own label is given once, with the theorems it rests on (Section 2.9): its if-and-only-if is shown in an idealized setting, for path laws over powers of one fixed operator, and bounds for realized products of distinct layers and the Dobrushin bound of Section 5.5 carry its direction, that mixing without short paths forgets, to stacks whose operators change from layer to layer. The Dobrushin bound requires strict mixing with shared feature maps, and these bounds do not model the full block. Its removal signatures (Section 5.2) were measured in trained networks, all below 25 million parameters.

Exact under premises checked network by network. The anchor exception, that capacity elsewhere cannot recover at bounded gain what short paths carry (Section 5.5): holds for any stack whose layer maps restrict to a common discriminative subspace with norms at most one and whose contraction product tends to zero, with the recovery gain uniformly bounded (a Dobrushin version covers sink-like attention with shared feature maps); these premises are checked layer by layer.

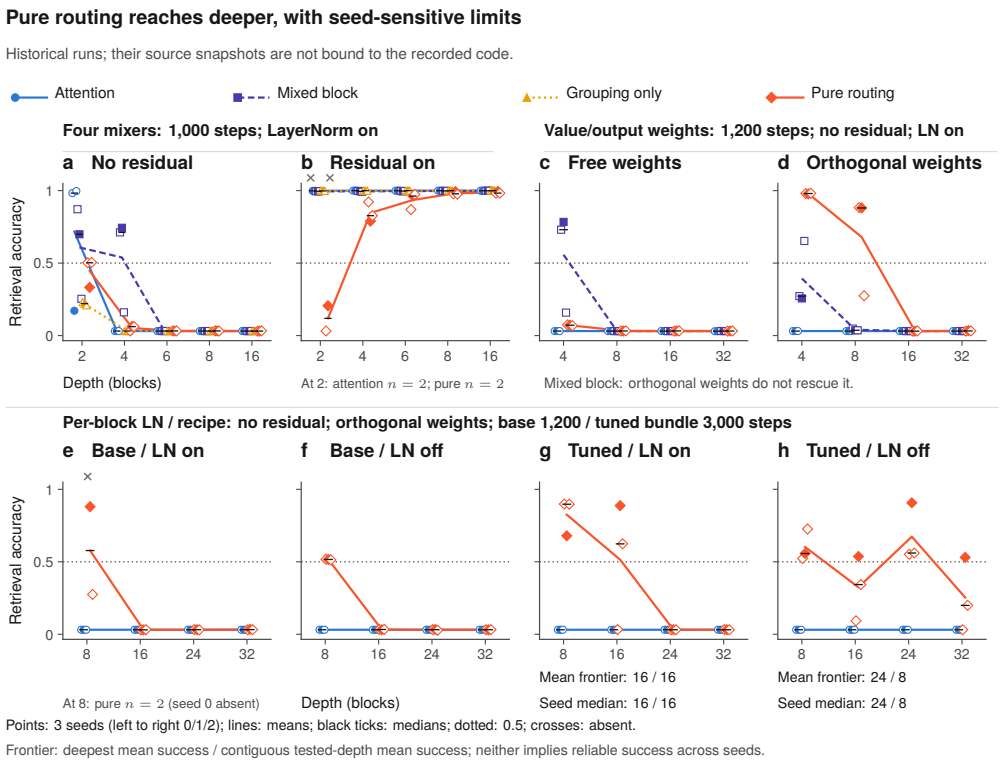


Figure 5.3: **Orthogonal pure routing reaches deeper without a residual, but outcomes depend on seeds and the training recipe.** (a–b) The first study: four mixers with and without residual connections, namely attention, pure routing, grouping only, and a mixed block that averages grouping and conservative routing. (c–d) The second study constrains the value and output weights: pure routing improves, while the mixed-block hypothesis stated in advance remains unsupported (its test was inconclusive). (e–h) The third study varies per-block LayerNorm (LN) and the recipe; the final LayerNorm remains on. All 285 training records are shown: 118, 72 and 95 from the three studies in turn. Three seed slots (0/1/2) are slightly offset within each arm; seed 1 is filled. Coincident marks overlap. Lines connect seed means at equally spaced tested depths, black ticks mark medians, and the dotted threshold is 0.5. These are seed observations, not confidence intervals. Crosses above the axes mark absent records: in the first study, residual-on depth 2, attention seed 0 and pure-routing seed 2; in the third, base/LN-on depth 8, pure-routing seed 0. Those means use $n = 2$; no value is imputed. The synthetic task copies a partner's content among 32 tokens (chance 1/32), evaluated on 512 examples. Later snapshots of the code specify width 64, batch 128 and Adam at base learning rate 0.003. The third study's tuning bundles identity-initialized orthogonal weights, a zero-initialized rotation gate for pure routing only, warmup and cosine scheduling, and 3,000 rather than 1,200 updates. Budgets and initialization differ across studies; these are descriptive comparisons, not isolated causal effects. The frontier recorded in the third study is the deepest tested depth with mean accuracy at least 0.5; the stricter contiguous rule also requires every shallower tested depth to pass. Tuned pure-routing seed frontiers (deepest/contiguous) are $(8, 16, 16)/(8, 16, 16)$ with LN on and $(24, 32, 24)/(8, 32, 8)$ with LN off; their medians and the mean-based summaries are labelled. Six deterministic reversibility checks are separate, not training trials. The code snapshots cannot be matched to the code that produced the runs, and no replication is inferred.

Chapter 6

A free content path: alignment and separation

Chapter 5 kept short paths, so that content survives depth. The second condition concerns what travels along them. Chapter 0 states it in brief (Section 0.3.2): deciding which tokens to read and deciding what to carry from them are different jobs. A head can keep them apart only if its content map can be chosen without constraining its structure map. Section 6.1 states when an independent content map is sufficient and when it is necessary; Section 6.2 shows that applying gain before or after a second map leaves one of two channels open; and Section 6.3 prices the violation in a router whose one channel carries both the address and the cargo. Section 6.4 turns from content to blocks of tokens: a leak across a block boundary can grow with path length, while exact zeros stop it. Unlike the anchor, this condition can be bought back by capacity elsewhere (Section 6.6), and Section 6.7 says what of it holds in trained networks.

6.1 The separation theorem

An architecture does not choose where gain sits relative to a second map; it chooses a parameterization, and training chooses the parameters. Which parameterization lets training keep nuisance out of the content a head transports? A *propagation architecture* is a pair $(\mathcal{G}_\Phi, \Pi_\Psi)$: a structure map $\mathcal{G}_\Phi$ that builds a row-stochastic operator from the data, and a content map Π_Ψ, applied row by row, that selects what is transported. One hop is $F(X) = \mathcal{G}_\Phi(X)\, X\, \Pi_\Psi$; hops are aggregated by an amplifier, optionally beside an identity path. For a fixed reversible operator with eigenpairs (μ_j, v_j) and a content map $\Pi_\Psi = \sum_k \sigma_k r_k s_k^\top$, the amplified hop maps the tensor mode $v_j r_k^\top$ to $g(\mu_j)\,\sigma_k\, v_j s_k^\top$. Every propagated component is thus weighted by the product of a graph gain and a content gain, and either factor at zero removes it. The operator cannot avoid amplifying nuisance *patterns*, which may be among its slowest modes, but the content map can refuse to carry nuisance *content*. It can do so for any operator when it can be chosen without constraining the operator:

Definition 6.1 (Structure–content separation)**.** $(\mathcal{G}_\Phi, \Pi_\Psi)$ is *separated* if (S1) $\mathcal{G}_\Phi$ depends on Φ alone, (S2) Π_Ψ depends on Ψ alone, and (S3) the parameter space factorizes, $\Theta = \Theta_\Phi \times \Theta_\Psi$, with no constraint coupling the factors.

Standard attention is separated, with $\Phi = \{W_Q, W_K\}$ and $\Psi = \{W_V\}$. Tying $W_V = W_K$ violates (S2), and tying W_V to any function of the structure parameters violates (S3).

Sufficiency. Let the features split as $U_{\text{disc}} \oplus U_{\text{nui}}$. A separated architecture can choose Π_Ψ with range in U_{disc} whatever Φ is, provided its content family contains the projection onto U_{disc}.

The transported content then has no nuisance coordinates, whatever sample pattern the operator amplifies, and with a nonnegative normalized aggregator every output row lies in the convex hull of discriminative content. What remains is subtractive: a nuisance-aligned operator averages good content within the wrong clusters, which shrinks it but never injects nuisance, and the identity path keeps the unpropagated content.

Necessity holds for an identified model. A spectral gap of a kernel operator does not certify its cluster structure, so it cannot carry the converse: for one Gaussian cloud with no classes at all, the population row-normalized Gaussian kernel is a Gaussian autoregressive chain, whose mixing gap is $c^2/(1+c^2)$ for bandwidth c in units of the data scale.

Remark 6.2 (A gap without classes). Let $x \sim \mathcal{N}(0, \sigma^2)$ and $k(x, y) = \exp(-(x-y)^2/(2h^2))$ with $\sigma, h > 0$. Completing the square in y shows that the population row-normalized kernel $P(x, \mathrm{d}y) \propto k(x, y)\,\mathcal{N}(y; 0, \sigma^2)\,\mathrm{d}y$ is the Gaussian $\mathcal{N}(ax, v)$ with $a = \sigma^2/(\sigma^2+h^2)$ and $v = \sigma^2 h^2/(\sigma^2+h^2)$. This autoregressive chain is reversible with respect to its stationary Gaussian law $\mathcal{N}(0, v/(1-a^2))$, its eigenfunctions are the Hermite polynomials of that law, and its eigenvalues are a^k, $k \geq 0$. Its gap is $1-a = c^2/(1+c^2)$ with $c = h/\sigma$: positive, although the data have no cluster structure at all.

A positive gap does not identify labels. Necessity therefore needs a model in which class and nuisance are identified. Let Q_z and Q_y average over nuisance and class labels on a balanced product population, so that they commute with each other and with the uniform average U, and let

$$P_0 = (1-a-b)\,U + a\,Q_z + b\,Q_y, \qquad a, b \geq 0, \quad a+b \leq 1.$$

Its centered nuisance space Z has eigenvalue a, its centered class space eigenvalue b, and interactions eigenvalue zero.

Theorem 6.3 (Separation necessity on an identified block model)**.** *Let $\widehat{P} = P_0 + E$ be symmetric and stochastic with $\|E\|_2 \leq e$ and $a - b > 2e$. The leading $d_z = \dim Z$ centered eigenspace $\widehat{V}$ of $\widehat{P}$ satisfies*

$$\|(I - \Pi_Z)\widehat{V}\|_F \leq \frac{\sqrt{d_z}\,e}{a-b-e}.$$

If a nuisance resource $w \geq 0$ of the structure map obeys $a \leq Cw$ and a leading centered eigenvalue is at least $t > e$, then $w \geq (t-e)/C$. For $E = 0$, a filter with nonnegative gains g and $g(b) > 0$ multiplies the nuisance-to-class amplitude ratio of equally normalized inputs by $g(a)/g(b)$.

The proof is in Section A.7. The theorem is conditional on the block identification, the resource relation $a \leq Cw$ and the error event $\|E\|_2 \leq e$, and each must be established in a given data model. For Gaussian clusters two further results supply counterparts of all three for an identified four-group model, provided the nuisance eigenvalue of the noisy kernel is matched to its ideal value, for instance by an isolated ideal band. A uniform finite-sample comparison bounds the distance between the normalized kernel of noisy data and that of its centers, simultaneously over every linear metric of bounded norm and every bandwidth above a floor, including ones selected from the same sample:

Theorem 6.4 (Uniform finite-sample normalized-kernel comparison)**.** *Let $x_i = c_{g_i} + \sigma z_i \in \mathbb{R}^d$ with standard Gaussian z_i, fixed group labels g_i and centers of diameter at most D. For every linear map A with $\|A\|_2 \leq B$ and bandwidth $h \geq h_{\min} > 0$, let $W_{ij} = e^{-\|A(x_i - x_j)\|^2/(2h^2)}$ and $W^0_{ij} = e^{-\|A(c_{g_i} - c_{g_j})\|^2/(2h^2)}$, self-loops included, with degree matrices D_W, D_0 and normalized kernels $H = D_W^{-1/2} W D_W^{-1/2}$, $H_0 = D_0^{-1/2} W^0 D_0^{-1/2}$. For $0 < \delta < 1$ put $R_\delta = \sqrt{d} + \sqrt{2\log(n/\delta)}$, $\nu = B\sigma R_\delta$, $a' = 2(BD\nu + \nu^2)/h^2_{\min}$ and $e = e^{2a'} - 1$. With probability at least $1-\delta$, simultaneously for every allowed A and h, $e^{-a'} W^0_{ij} \leq W_{ij} \leq e^{a'} W^0_{ij}$, the same holds for the degrees, and $\|H - H_0\|_2 \leq e$; each ordered eigenvalue of H is within e of that of H_0.*

Here a' grows with the mapped noise radius relative to the bandwidth, so the bound is informative only at high signal-to-noise ratio. On four equally sized groups the ideal operator is explicit:

Proposition 6.5 (Identified four-group model)**.** *Take equal numbers of points at the centers $c_{yz} = y\,m_y u + z\,m_z v$, $y, z \in \{\pm 1\}$, with $m_y, m_z > 0$ and orthonormal u, v, and a metric with $\langle Au, Av\rangle = 0$, $\|Au\|^2 = w_y$ and $\|Av\|^2 = w_z$. The ideal operator H_0 has group-level eigenvalues 1, $\theta_y = \tanh(w_y m_y^2/h^2)$, $\theta_z = \tanh(w_z m_z^2/h^2)$ and $\theta_y\theta_z$, on the constant, class, nuisance and interaction functions; all within-group contrasts have eigenvalue zero.*

The proofs of Theorem 6.4 and Proposition 6.5 are in Section A.7. Nuisance leads exactly when $w_z m_z^2 > w_y m_y^2$: when the metric makes the nuisance the cleaner clustering, as backgrounds, styles and acquisition artefacts can be. If the identified nuisance eigenvalue of the noisy operator is at least $t > e$, eigenvalue perturbation gives $w_z \geq (h^2/m_z^2)\operatorname{artanh}(t - e)$ for $0 < t - e < 1$, provided the correspondence to θ_z is established, for instance by an isolated ideal band. Read together, these say that on the identified model a structure map whose leading modes are nuisance-dominated must spend nuisance resource, and a filter with $g(b) > 0$ then amplifies the nuisance component relative to class by $g(a)/g(b)$. A tied content map that inherits the structure map's nuisance weight transports that component; a separated one can decline it. For couplings whose content inherits the structure map's nuisance weight, as exact tying does, two independently chosen parameter groups are therefore the minimum for this model; a coupling that projects that weight away, such as one whose content map excludes the structure map's range, is not covered. We do not claim necessity for unrestricted learned metrics, which remains open.

Range containment and dimensional capacity. Two elementary facts locate the conflict inside one head. Write the content block of W_K as $E \in \mathbb{R}^{r \times d_b}$, for content of dimension d_b and head width r.

Proposition 6.6 (Range containment and dimensional capacity)**.** *(i) If $W_V = MW_K$, the output of a head at token i is $M\sum_j P_{ij}k_j$, and $\ker W_K \subseteq \ker W_V$. (ii) Let $E \in \mathbb{R}^{r\times d_b}$ be the content block of W_K. If content is recovered injectively from the keys, then $r \geq d_b$, and at $r = d_b$ every nonzero query q has $E^\top q \neq 0$. (iii) For a query span of dimension s, there is an injective E with $E^\top q = 0$ for every query q in the span if and only if $d_b \leq r - s$.*

The proof is in Section A.7. By (i), a head whose value map factors through the key map carries nothing the keys do not encode. By (ii), injective content recovery from the keys forces $r \geq d_b$, and at $r = d_b$ the content block is invertible, so every nonzero query reads content; by (iii), injective content storage that no query sees is possible if and only if $d_b \leq r - s$. A tied head at the content width therefore cannot route blind to content: the logits that decide where to look move with the content being carried. Invisibility that is only approximate has a price too: with a decoder of gain at most L_D, the content's exposure to a query metric G is at least L_D^{-2} times the sum of the d_b smallest eigenvalues of G.

Proposition 6.7 (Stable recovery prices approximate invisibility)**.** *Let $G \succeq 0$ be an $r\times r$ query metric with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_r$, and let $DE = I_{d_b}$ with $\|D\|_2 \leq L_D$. Then $\operatorname{tr}(E^\top G E) \geq L_D^{-2}\sum_{j=1}^{d_b}\lambda_j$, and the bound is attained.*

The proof is in Section A.7.

A restricted capacity theorem. The routing budget of Chapter 9 and the content fluctuations it must dominate can be priced together. Fix one query q with $\|q\| \leq Q$, a target $*$ and $n-1$ competitors, keys $k_j = a_j + Eb_j + \xi_j$ with address margins $q^\top(a_* - a_j) \geq \Delta$, content $\|b_j\| \leq B$, key noise $\|\xi_j\| \leq \nu$, and logits $q^\top k_j/\tau$. Put $e_q = \|E^\top q\|$, the query's exposure to content.

Theorem 6.8 (Robust routing and value recovery)**.** *With $m = \Delta - 2Be_q - 2Q\nu$, the target's softmax share satisfies $p_* \geq (1 + (n-1)e^{-m/\tau})^{-1}$. If the transported values obey $\|v_j - b_j\| \leq \nu_v$, the output error is at most $2B(1-p_*) + \nu_v$. The budget*

$$\Delta \;\geq\; 2Be_q + 2Q\nu + \tau \log \frac{(n-1)(1-\eta)}{\eta}$$

suffices for $p_* \geq 1 - \eta$, and it is necessary for a uniform guarantee over the subclass with equal address gaps, freely chosen bounded content and noise, and $\|q\| = Q$.

The proof is in Section A.7. The theorem prices address margin, content scale, noise, width (through e_q) and readout gain (through ν_v, at most $L_D\nu$ for a tied decoder D of gain L_D satisfying $Da_j = 0$ and $DE = I$) separately. It is a restricted capacity theorem, not a Bayes-risk lower bound for learned metrics. For a linear Gaussian key channel with power and query-exposure budgets the optimal reconstruction error over all measurable decoders is explicit, and exact query blindness costs width: only $\min(d_b, r - s)$ content directions can be stored.

Theorem 6.9 (Optimal Gaussian reconstruction with two budgets)**.** *Let $b \sim \mathcal{N}(0, \tau_b^2 I_d)$ and $k = Eb + \zeta$ with $\zeta \sim \mathcal{N}(0, \sigma_k^2 I_r)$ independent, and impose $\operatorname{tr}(E^\top E) \leq P$ and $\operatorname{tr}(E^\top G E) \leq \varepsilon^2$ for a query metric $G \succeq 0$ with eigenvalues $g_1 \leq \cdots \leq g_r$. With $k_0 = \min(d, r)$, the smallest expected squared error over all measurable decoders is*

$$(d - k_0)\tau_b^2 + \min\Big\{ \sum_{j=1}^{k_0} \frac{\sigma_k^2}{\sigma_k^2/\tau_b^2 + p_j} \;:\; p_j \geq 0,\ \sum_j p_j \leq P,\ \sum_j g_j p_j \leq \varepsilon^2 \Big\},$$

attained by encoding along the weakest query eigenvectors and decoding with the posterior mean. If the queries span s dimensions and must see no content, the optimum under $\|E\|_F^2 \leq P$ is $(d-k)\tau_b^2 + k\tau_b^2/(1 + \tau_b^2 P/(k\sigma_k^2))$ with $k = \min(d, r - s)$, and $d\tau_b^2$ when $k = 0$.

The proof is in Section A.7.

Evidence. On a synthetic family built so that the nuisance is the cleaner clustering (an estimator-level check in closed form, with a relative output noise floor and without gradient training; means over three data draws) , the tied architecture's excess risk, when its amplification is engaged, tracks the dynamic range of Section 4.3 (Spearman 1.0; +13 pp over the sweep; left free, it keeps its realized range below 2.7 and stays flat, +0.6 pp) while the separated and benign arms stay flat. In an oracle construction on a synthetic routing-and-content task (fixed routing and value maps and a least-squares readout, with no training), a head whose value range is contained in its key range, and which pays its routing cost out of the content dimensions at any width, stays at risk 0.311–0.312 at every width from the content dimension up (0.768 at half of it), while an untied head and a single-group coupling that excludes the key range reach the noise floor 0.079 at those widths (means over five seeds). The failure is dimensional in trained one-hop attention models (five seeds at each width and token count): the tied map fails at head width $\leq d_b$ and recovers above it, with value noise, retrieval rising $0.29 \to 0.96 \to 1.00$ from d_b to $2d_b$ to full width (separate maps reach 0.97 at d_b) while an alignment witness falls $1.0 \to 0.22 \to 0.01$ (at $2d_b$, 0.02 with 32 tokens and 0.42 with 128; means over both) as the tied map splits its output into a routing part and an orthogonal content carrier. Trained networks buy this evasion wherever width allows (Section 6.6).

What separation does not buy. Separation needs one structure group and one content group, not a split of the structure group into queries and keys. A single symmetric structure map satisfies (S1) equally well. The query–key split buys asymmetric scores, and with them circulation, which is the arrow (Chapter 10), not safety. Nor does a feature-side map change the graph gain ratio between the constant and other sample modes; that is the anchor's job (Chapter 5).

6.2 The commutator of the two orders

Section 4.3 set an amplifier $g(P)$, applied to a seed Y (labels or features, one column per channel), beside a second linear map M on the token index: a second smoothing kernel, a projection onto a subspace of node functions, the token-mixing part of a readout. The two orders are

$$Y_{\text{post}} = g(P)\,MY \quad \text{(project, then amplify)}, \qquad Y_{\text{pre}} = M\,g(P)\,Y \quad \text{(amplify, then project)},$$

and practice usually fixes the order without comment. Only maps on the token index raise the question: a map on the feature index, $X \mapsto XW$, commutes with $g(P)$ by associativity.

Lemma 6.10 (Ordering identity)**.** $Y_{\text{pre}} - Y_{\text{post}} = [M, g(P)]\,Y$. *If P is reversible with eigenpairs (μ_k, v_k) and $k_{jl} = \langle v_j,\, Mv_l\rangle_\pi$, then*

$$\langle v_j,\, [M, g(P)]\,v_l\rangle_\pi \;=\; k_{jl}\,(g(\mu_l) - g(\mu_j)).$$

If moreover M is π-self-adjoint, $[M, g(P)]$ is π-antisymmetric, so $\langle Y,\, Y_{\text{pre}}\rangle_\pi = \langle Y,\, Y_{\text{post}}\rangle_\pi$ for every real Y.

Proof. The first claim is the definition of the commutator. For the second, $g(P)v_l = g(\mu_l)v_l$ and, $g(P)$ being π-self-adjoint, $\langle v_j,\, g(P)w\rangle_\pi = g(\mu_j)\langle v_j,\, w\rangle_\pi$ (Section A.1, T1); subtract. For the third, $[M, g(P)]^\dagger = [g(P), M] = -[M, g(P)]$ when both are self-adjoint, and a real antisymmetric quadratic form vanishes (Lemma 1.2(i)). □

Every ordering error is therefore a product of two factors: a *coupling* k_{jl}, how much the second map mixes the amplifier's modes, and a *gain difference* $g(\mu_l) - g(\mu_j)$, how differently the amplifier treats them. If the map reads in the amplifier's own frame ($k_{jl} = 0$ for $j \neq l$), or the gain is flat across the modes it couples, the orders agree. The last clause is the bridge to Section 3.5: like the flow J, the commutator of an amplifier with a self-adjoint map is antisymmetric in the walk's geometry, so the two orders agree on the seed's own direction and differ only off it, where cross readouts can see the difference. How large the difference can be is the fragility identity (Theorem 4.1).

An earlier account read a separated head as the architectural form of "project, then amplify". The fragility identity shows that neither order dominates: amplifying before a second map exposes the escaped part of the signal to whatever that map reads, and amplifying after it leaves in the output a component a trusted projection would have removed (Theorem 4.1(iii)–(iv)). A tied head builds its content map from the representation that carries the nuisance, which is the first channel. Separation closes that channel for nuisance that rides in the features. It leaves the second channel open, and the within-cluster averaging of Section 6.1; later maps, and the anchor, which caps gain ratios on a lazy reversible operator (Corollary 4.2), are what act on those.

6.3 Address and cargo

A routing machine carries two things: an address, which decides who talks to whom, and a cargo, which is what is said. They must travel in separate channels. If one channel carries both, the cargo becomes noise on the address. The general statement is the separation theorem of Section 6.1. Here we price the simplest violation, in the model that a constructed separation test uses (Bruce, 2026e). One channel carries both the address and the cargo: item j of class y_j scores $(z_j + g_{y_j})/2$, where z_j is its address's cosine with the query and g_{y_j} is the query's cosine with the cargo of class y_j. The address therefore arrives at half its amplitude. *Soft routers lose gain.* For softmax, with weights proportional to $e^{\tau s_j}$ for scores s_j, halving the address amplitude is the same as halving τ, and the

cargo channel adds a background of its own. For a target of address score c, the mean-field gain wall of Section 9.3, $\tau c - \ln m(\tau)$ with $m(\tau) = \mathbb{E}e^{\tau z}$, falls to $\tau c/2 - 2\ln m(\tau/2)$. The separation test (isotropic keys, $D = 128$, exact queries) samples $\ln N$ in steps of 1.4 nats, and each mean-field onset falls inside the step where the measured success crosses one half. Its records are historical constructed-key records, not reproduced, and its onsets are read on this grid; on the recovery's finer grid they differ slightly. The tied router fails between $\ln N = 4.16$ and 5.55 at $\tau = 10$ and between 8.32 and 9.70 at $\tau = 20$, against mean-field onsets of 4.80 and 9.22. The untied router fails between 8.32 and 9.70 at $\tau = 10$, against 9.61, and at $\tau = 20$ not before the grid ends at $\ln N = 11.09$. At matched effective gain the tie costs only the cargo's background, $\ln m(\tau/2)$: 0.10 nats at $\tau = 10$. For a soft router, contamination is therefore mostly lost gain, which more gain can buy back where gain control allows (Section 7.6).

A hard router loses evidence. A router with no gain to lose takes the contamination in its evidence.

Theorem 6.11 (The contaminated argmax)**.** *Let the target's address score be c, let the other address scores be independent with continuous distribution function F, let C cargo scores $G_1, \dots, G_C$ be independent and independent of the addresses, and let each item belong to an independent, uniformly random class. Score item j by $(Z_j + G_{Y_j})/2$. With $F_g(x) = \frac{1}{C}\sum_{k=1}^{C} F(x - g_k)$, the target is the argmax of N items with probability*

$$p_N = \mathbb{E}_G\Big[\frac{1}{C}\sum_{y=1}^{C} F_G(c + G_y)^{N-1}\Big],$$

while without cargo it is $F(c)^{N-1}$. Unequal class probabilities replace both uniform averages by weighted sums.

The proof is in Section A.7. Cargo is shared by all items of a class. Replacing it with fresh noise per item is a different model and need not give the same cost. The formula is the exact cost of contamination for an argmax. It does not say that the contaminated argmax is the best decision from contaminated scores. In the separation test ($D = 128$, 16 classes, $c = 0.35$), a hard router fails at $\ln N = 10.27$ with clean scores and at 6.19 with tied ones, a ratio of 0.60. Walkers that select, which route by sampling (Section 9.3), with $M = 256$ or 65,536 samples at $\tau = 10$ or 20, show ratios of 0.60–0.67. Without cargo, the theorem gives $\ln N = 10.27$ in closed form; integrating its conditional law numerically gives 6.16 with cargo, a ratio of 0.59980, and the integrated law's largest gap from the measured tied curve is 0.019 in probability. That ratio is a numerical evaluation for this model, from six scrambled quasi-Monte Carlo replicates, not a universal constant or a rigorous enclosure (Section A.7).

Factorized addresses are immune. Separation can be built in. Give the addresses product structure, so that item (i, j) scores $a_i + b_j$.

Proposition 6.12 (Clean factorized addresses)**.** *The global argmaxes of $a_i + b_j$ are the pairs of half-argmaxes, and a Cartesian top-k shortlist built from the two clean half-scores contains one of them for every $k \geq 1$, and the global argmax itself when it is unique. If routing and final selection use only those scores, no change in the cargo changes the selection. If the final selection instead scores $a_i + b_j + v_{ij}$, the clean winner (i_*, j_*) stays the unique winner exactly when $v_{ij} - v_{i_*j_*} < a_{i_*} + b_{j_*} - a_i - b_j$ for every other shortlisted pair, and stays a winner exactly when these inequalities hold with $\leq$.*

The proof is in Section A.7. Immunity holds by construction of the channels. It does not make a noisy address find its target: inclusion in the shortlist still depends on the address's own

evidence, and Section A.7 gives its exact conditional law for random codebooks. On tied scores the separation test's flat hard router finds the target only 42–54% of the time at $c = 0.5$, less often with 64 codewords per half than with 32. In the recovery's readout of the same draws, admitted as a registered rebaseline, per-half search on clean codebooks finds the target 99.3% and 97.9% of the time with 32 and 64 codewords per half, at every tested shortlist size from $k = 1$ to 16: exactly as often as the clean global argmax is the target, and at $c = 1$ always. Rescoring the $k \times k$ shortlist with tied scores lets the cargo back in, and more of it as the shortlist grows: the target is found 72.3% and 66.2% of the time at $k = 2$, and 53.7% and 43.1% at $k = 16$, close to the flat router on tied scores (Bruce, 2026e).

6.4 Leaks can grow with path length; exact block separation stops them

Separation keeps two jobs apart inside one head. A partition of the tokens into blocks, such as the documents of one packed context, keeps groups of tokens apart, and a head respects it only as far as it puts no mass across a block boundary. Dense softmax puts positive mass on every admissible token (Section 2.5), so a dense head always puts some mass across a boundary, and a positive leak can grow with path length. Fix a partition of tokens.

Theorem 6.13 (Path-law leakage bound)**.** *Suppose P is stochastic and its probability of leaving the current block in any one step is at most $\epsilon \in [0, 1]$, uniformly over rows. The probability of ending outside the starting block after k steps is at most $1 - (1 - \epsilon)^k$. For a nonnegative filter $g(P) = \mathbb{E}P^T$, with an integer-valued finite-a.s. path length T independent of the chain, its cross-block row mass is at most*

$$1 - \mathbb{E}(1 - \epsilon)^T \le \epsilon \mathbb{E}T.$$

If $\mathbb{E}[T(T - 1)] < \infty$, then

$$0 \le \epsilon \mathbb{E}T - \{1 - \mathbb{E}(1 - \epsilon)^T\} \le \tfrac{1}{2}\epsilon^2 \mathbb{E}[T(T - 1)].$$

The proof is in Section A.7. The path length, the lottery of Section 1.3, is written T here because K will count blocks. An absorbing second block with an independent departure chance ϵ at every step attains $1 - (1 - \epsilon)^k$, so this is a sharp uniform bound. If $\mathbb{E}T < \infty$, dominated convergence gives slope $\mathbb{E}T$ at zero for this extremal example. If $\mathbb{E}T = \infty$, the slope is infinite by Fatou's lemma; there is no finite general first-order coefficient. Signed filters $g(P) = \sum_{k \ge 0} c_k P^k$ retain exact zeros, but lose the probability interpretation; a derivative bound then uses $\sum_k |c_k| k$ where finite, rather than a signed "mean". How far a leak saturates, and how fast it grows, depend on the leak metric and the normalization, and in a balanced block model both are exact.

Proposition 6.14 (Balanced block model)**.** *Let B average within each of $K \ge 2$ equal-size blocks and let $U = \mathbf{1}\mathbf{1}^\top/n$ average globally. For $P_\epsilon = (1 - \epsilon)B + \epsilon U$ and integer $L \ge 1$,*

$$P_\epsilon^L = U + a(B - U), \qquad a = (1 - \epsilon)^L.$$

If off *deletes entries inside the prescribed diagonal blocks, the normalized Frobenius contamination is exactly*

$$C_L(\epsilon) = \frac{\|\mathrm{off}(P_\epsilon^L)\|_F}{\|P_\epsilon^L\|_F} = \sqrt{1 - \frac{1}{K}}\,\frac{1 - a}{\sqrt{1 + (K - 1)a^2}}.$$

It tends to $\sqrt{1 - 1/K}$ as $a \to 0$, and $C_L'(0) = L\sqrt{K - 1}/K$. In particular $C_L(\epsilon)/C_1(\epsilon) \to L$ as $\epsilon \downarrow 0$.

The proof is in Section A.7. The one-step row probability of crossing in this model is $\epsilon(1-1/K)$, not ϵ, so the slope depends on which leak metric is recorded. For unequal block sizes n_b the limiting uniform-matrix saturation is $\sqrt{1-\sum_b (n_b/n)^2}$, and $\sqrt{1-1/K}$ is not a universal saturation for every filter, kernel, stationary distribution or normalization. Every fixed finite-power expression above is continuous in ϵ. The mathematical "cliff" is a steep crossover on a scale $1/L$, and the limits do not commute: at any fixed $\epsilon > 0$, $L \to \infty$ saturates; at exactly zero leak, every depth has zero contamination. For a random path law the relevant scale is its path distribution, with its mean controlling only the small-leak regime under the moment conditions above. The same block-averaging model gives the exact contamination of every filter, with unequal blocks and an identity atom allowed. Partition n coordinates into K nonempty blocks of sizes n_b, let B average uniformly within each block, $U = \mathbf{1}\mathbf{1}^\top/n$ and $P_\epsilon = (1-\epsilon)B + \epsilon U$, $0 \le \epsilon \le 1$; let off remove all entries within the diagonal blocks, and put $\xi = 1 - \sum_b (n_b/n)^2$.

Theorem 6.15 (Three spectral sectors give the exact contamination)**.** *For any real function g defined at $0, 1-\epsilon, 1$, write $h = g(1)$, $a = g(1-\epsilon)$, $z = g(0)$. If $g(P_\epsilon) \neq 0$, then*

$$g(P_\epsilon) = hU + a(B-U) + z(I-B), \qquad \frac{\|\mathrm{off}(g(P_\epsilon))\|_F}{\|g(P_\epsilon)\|_F} = \frac{\sqrt{\xi}\,|h-a|}{\sqrt{h^2 + (K-1)a^2 + (n-K)z^2}}.$$

Coincident eigenvalues at the endpoints are interpreted with the same value of g; the formula remains valid.

Write $C_g(\epsilon)$ for this contamination of $g(P_\epsilon)$, and $C_P(\epsilon)$ for that of the one-step filter $g(P) = P$.

Corollary 6.16 (Path laws, exact slope and saturation)**.** *Let T be a finite-almost-sure nonnegative integer path length, independent of the chain, and $g(P) = \mathbb{E}P^T$. Write $p_0 = \mathbb{P}(T = 0)$ and $A(\epsilon) = \mathbb{E}(1-\epsilon)^T$, with $0^0 = 1$. Then exactly*

$$C_g(\epsilon) = \sqrt{\xi}\,\frac{1 - A(\epsilon)}{\sqrt{1 + (K-1)A(\epsilon)^2 + (n-K)p_0^2}}.$$

If $\mathbb{E}T < \infty$, its right derivative at zero is

$$C_g'(0) = \frac{\sqrt{\xi}\,\mathbb{E}T}{\sqrt{K + (n-K)p_0^2}}.$$

At full mixing,

$$C_g(1) = \sqrt{\xi}\,\frac{1-p_0}{\sqrt{1 + (n-1)p_0^2}}.$$

For $K \ge 2$, comparison with the one-step filter gives

$$\lim_{\epsilon \downarrow 0} \frac{C_g(\epsilon)}{C_P(\epsilon)} = \mathbb{E}T\sqrt{\frac{K}{K + (n-K)p_0^2}}.$$

The proofs of Theorem 6.15 and Corollary 6.16 are in Section A.7. For complex filters the three squared coefficients become squared moduli, and a common nonzero scaling of the filter cancels in the fraction. An identity-only path law has zero contamination for every leak. A power P^L, $L \ge 1$, has $p_0 = 0$ and $\mathbb{E}T = L$; the exponential $e^{t(P-I)}$ has Poisson path length, $p_0 = e^{-t}$ and mean t; the normalized resolvent $(1-\alpha)(I-\alpha P)^{-1}$ has $p_0 = 1-\alpha$ and mean $\alpha/(1-\alpha)$. For row i in block b, the *row-mass* leakage is instead $(1-A)(1-n_b/n)$, with slope $(1-n_b/n)\mathbb{E}T$. An identity atom

therefore changes the Frobenius denominator but not that row-mass slope factor, which is why the two metrics have different amplification ratios.

Exact zeros are the other side of the cliff. For block-diagonal P_0, whose blocks may hold arbitrary internal entries, $\mathrm{off}(g(P_0)) = 0$ for every analytic filter g defined at P_0, whatever the signs of its coefficients: exact block zeros keep every filter block-diagonal, at any depth and whatever the flow inside the blocks. The block-averaging model shows both sides at once. The contamination of Theorem 6.15 is proportional to $|h - a| = |g(1) - g(1-\epsilon)|$, which vanishes at $\epsilon = 0$ for every filter and is positive whenever a leak makes $g(1-\epsilon)$ differ from $g(1)$. Chapter 0 states the contrast in brief (Section 0.4.4).

Exactness also buys a freedom. Because the guarantee holds for every filter, the blocks of a block-diagonal operator need not share one: each block can be given its own filter and the operator stays block-diagonal, so no block's filter reaches another. What such per-block filters recover in practice is left for later work.

Cleaning is not carrying. Cleaning removes what should not be there, a nuisance direction or a leak across a boundary: it is subtraction. Carrying delivers content to a reader: it is addition, and it needs a channel with room for that content (Proposition 6.6). The records of Section 6.1 show both sides for one pair of heads: the tied head takes contamination that the separated head does not, and, in the trained one-hop models with value noise, it regains retrieval only as width makes room for a content carrier, from 0.29 at head width d_b to 0.96 at $2d_b$ (means over five seeds at each token count, pooled over 32 and 128 tokens). A direct comparison of cleaning and carrying is left for later work.

Figure 6.1 draws the two sides. Exact block zeros survive every defined analytic filter, while a small positive leak can accumulate with depth. In an exact construction with four blocks of two tokens, a one-step cross-block row mass ℓ grows to $\frac{3}{4}[1 - (1 - \frac{4}{3}\ell)^L]$ after L steps, the row-mass leakage above for the power P_ϵ^L: a 1% leak reaches 26.2% after 32 steps, while an exact zero stays zero at every depth. The construction is deterministic and makes no trained-model claim.

6.5 Leaks in trained and pretrained models

Pending: Learned and discovered blocks in trained models: cross-segment influence and accuracy against depth and the number of segments, for dense, masked, sliding-window, local, NSA-style and learned-block attention.
Pending: Document masks on real tasks: accuracy, or exact match and F1, and the measured cross-document leak of each attention variant on multi-document question answering, long-document classification and packed pretraining.
Pending: How much pretrained language models leak across documents: cross-document attention mass per head and composed influence across depth, in public language models on packed sequences, with and without the sink columns.

6.6 The tying tax and its buy-back

Necessity theorems bound what one constrained map does in one application, and they are exact on their own constructions: for C classes, a rank-$(C-2)$ content map on the proof's construction (Section 15.1) sits at 90.16% against a predicted ceiling of 90.0%. Trained systems evade them by composition wherever another resource has room. A content map of the same rank, trained jointly with its readout, reaches 99.84%. Spare head dimensions let a tied value map split its output into a routing part and a content carrier (Section 6.1). At depth 24 and width 192 on CIFAR-100, MLP

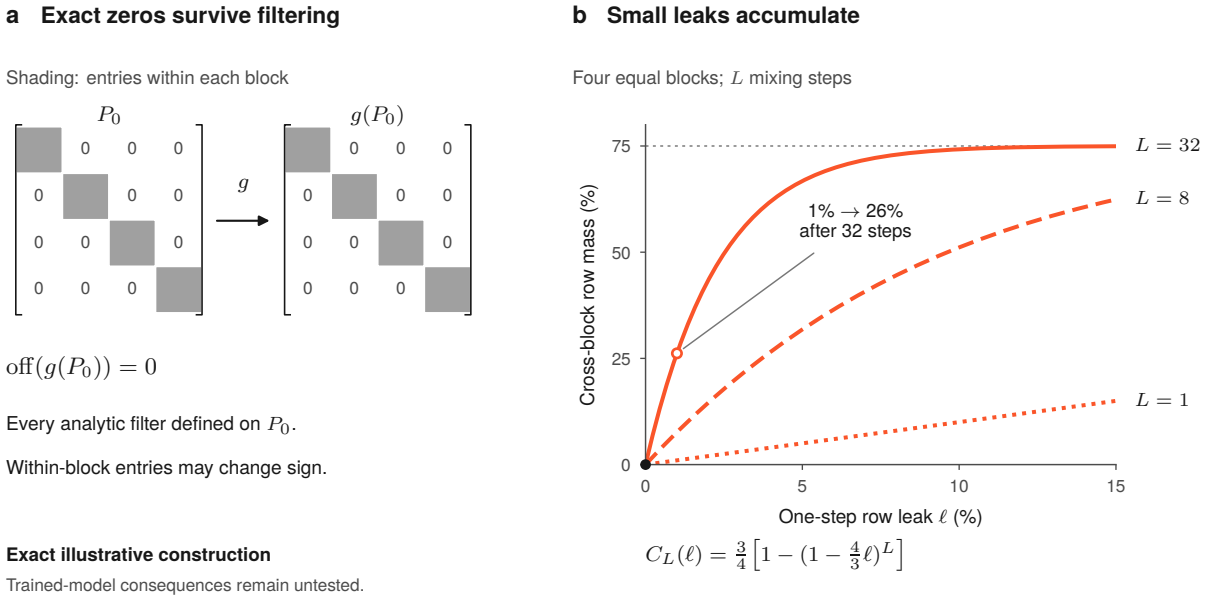


Figure 6.1: Exact block zeros survive every defined analytic filter, whereas a small positive leak can accumulate with depth. (a) For block-diagonal P_0, shaded blocks allow arbitrary internal entries, but $\mathrm{off}(g(P_0)) = 0$; off removes the within-block entries. (b) Exact illustrative construction: B averages within four blocks of two tokens and $U = \mathbf{1}\mathbf{1}^\top/8$ averages globally. For $P_\epsilon = (1-\epsilon)B + \epsilon U$, the one-step cross-block row mass is $\ell = 3\epsilon/4$, and $C_L(\ell) = \frac{3}{4}[1 - (1 - \frac{4}{3}\ell)^L]$ after L steps; here C_L is the cross-block row mass, not the Frobenius contamination of Proposition 6.14. Dotted, dashed and solid curves show $L = 1, 8, 32$; the grey line is the 75% row-mass limit as $L \to \infty$ at fixed $\ell > 0$. A 1% one-step leak gives 26.2% after 32 steps; exact zero stays zero at every depth. Finite-depth curves are continuous. The formula uses fractions and the axes show percentages. This deterministic construction has no sampling uncertainty and makes no trained-model claim.

capacity turns the value-tying tax from +1.70 to −0.72 pp as the MLP ratio grows from 0 to 4 (Section 0.3.4). Heads buy back subtraction: in a single small block, softmax difference retrieval rises from $R^2 = 0.46$ with one head to 0.998 with four. These are measurements of trained models. The neighbourhood of the standard design is therefore flat. The anchor is the exception, which Section 5.5 states with its premises.

The buy-back has a currency, and for one family of small vision transformers it has been priced. The tax is untied minus tied top-1 accuracy on CIFAR-100, in percentage points. More MLP capacity reduces it at depths 12 and 24: the interaction between tying and MLP ratio r is negative in all four configurations there, at −1.20 to −1.47 points per unit of $\ln(1 + r)$ (Holm-adjusted $p \leq 8.4 \times 10^{-6}$). The depth-4 mean tax stays positive over the measured range: at MLP ratio 4 it is +0.65 and +0.78 points at widths 96 and 192. At depth 24 and width 192 the tax is +1.70, +0.04 and −0.72 points at MLP ratios 0, 1 and 4, and the budget, the MLP ratio at which the interpolated mean tax first crosses zero, is $r^* = 1.11$, with a stratified 95% bootstrap interval of $[0.93, 1.63]$. The comparison is not parameter-matched: tied models have 21.6–24.6%, 15.1–16.4% and 7.9–8.3% fewer parameters at ratios 0, 1 and 4. The analysis covers 180 runs in 18 configurations, with five initialization-only seeds per arm that share one data order, and every interval is over those seeds. It is a reanalysis of a known outcome: it describes compensation at the level of accuracy, without a parameter-matched, mechanism or at-scale claim. Whether the budget holds at scale is the question

of Section 15.6.

6.7 What holds in trained networks

Exact under premises checked network by network. A free content path, necessary on an identified model when tied values inherit the keys' nuisance and bought back by capacity elsewhere (Sections 6.1 and 6.6): exact on the identified block model where necessity is stated; whether a trained network sits in that model must be checked, and with spare capacity trained networks evade it (compensation).

Shown in an idealized setting. The growth of leaks with path length, and the exactness of block zeros (Section 6.4): the growth bound is stated for path laws of one operator with a per-row crossing probability; exact block zeros stay zero for any stack whose layers share the blocks.

Measured in trained networks. Compensation, by which trained networks evade necessity (Section 6.6). A mechanism measured in trained networks: necessity theorems bind exactly on their constructions, and trained networks evade them when a resource is spare.

Chapter 7

Gain below the anchor

Feedback is the third failure mode. Attention's weights are computed from its own outputs, so it can reinforce itself: a change in the states changes the weights that average them, and the changed weights act on the states again. The condition that keeps this in check is gain control, which Chapter 0 states in brief (Section 0.3.2): hold the feedback below the anchor. The iteration that isolates the loop is the one of Section 1.4, $X_{t+1} = \alpha\mathcal{P}(X_t)X_t + \beta A$, one builder applied again and again with a share β of the anchors retained at every step; its feedback constant $\kappa_{\mathrm{f}} = RL_P$ and its margin $m = \beta - \alpha\kappa_{\mathrm{f}}$ are those of (1.5). Section 7.1 proves that a positive margin certifies a unique attracting state and that the threshold cannot be improved over the class; Section 7.2 computes softmax's feedback constant; Sections 7.3 and 7.4 take weight decay as the anchor and correct the first test of that reading; Section 7.5 asks what normalization and logit control certify; Section 7.6 shows that what gain control buys is uniqueness; and Section 7.8 says what of this holds in trained networks.

7.1 The certificate

On the ball B_R of Section 1.4 the builder has Lipschitz constant L_P, the feedback constant is $\kappa_{\mathrm{f}} = RL_P$, and the margin is $m = \beta - \alpha\kappa_{\mathrm{f}}$ (1.5). A positive margin is enough for convergence, and the threshold is sharp over the class of smooth builders (Bruce, 2026h).

Theorem 7.1 (Convergence and class sharpness)**.** *The ball B_R is invariant under (1.4). If $m > 0$, the update is a contraction with upper factor $q = \alpha(1 + \kappa_{\mathrm{f}}) = 1 - m < 1$, and every initialization in B_R converges geometrically to its unique fixed point. For every $\alpha \in (0, 1)$ and $\varepsilon > 0$, there are a smooth row-stochastic builder with $N = 2$, $d = 1$, $R = 1$, $\beta < \alpha L_P \le (1 + \varepsilon)\beta$, and an anchor with $\|A\|_{\infty,2} = 1$ such that (1.4) has at least two locally attracting fixed points in B_1.*

The proof is in Section A.8.

Values and weights contribute separately. Split the difference of two updates into $\alpha\mathcal{P}(X)(X - Y)$ and $\alpha(\mathcal{P}(X) - \mathcal{P}(Y))Y$. Their norms are bounded by $\alpha\|X - Y\|_{\infty,2}$ and $\alpha RL_P\|X - Y\|_{\infty,2}$, respectively, and together they give the upper factor $q = \alpha(1 + \kappa_{\mathrm{f}})$. Banach's theorem then gives convergence.

Why is the threshold sharp? Two scalar states suffice. A smooth mixer can make their contrast map cross the identity three times, with two attracting roots and a repelling root between them, while αL_P approaches β from above. Their mean contracts independently. This is sharpness over the builder class, not a claim that every mixer beyond the boundary fails. The photographic denoiser of Section 12.3, with the Gaussian NLM kernel (4.1), illustrates distinct attracting states; it is not

a kernel-specific sharpness result. Below the threshold, the margin also bounds how far the fixed point moves with its anchors.

Corollary 7.2. *If $m > 0$, then on a common invariant ball two anchors satisfy*

$$\|X^*(A') - X^*(A)\|_{\infty,2} \le \frac{\beta}{m}\|A' - A\|_{\infty,2}. \tag{7.1}$$

A candidate in that ball has solution error at most its update residual divided by m.

Both conclusions require a valid domain bound, not a sampled derivative or an observed convergence rate. The proof is in Section A.8.

The share need not be one number. Let node i mix with its own share, which may depend on the state,

$$F(X)_i = a_i(X)\,(\mathcal{P}(X)X)_i + (1 - a_i(X))A_i, \qquad 0 \le a_i(X) \le b \le 1 \text{ on } B_R.$$

Proposition 7.3 (Shares that depend on the state)**.** *Let $\|A\|_{\infty,2} \le R$, let the builder be row-stochastic and L_P-Lipschitz on B_R, let each share be S-Lipschitz there, and let $\|\mathcal{P}(Y)Y - A\|_{\infty,2} \le D$ for every $Y \in B_R$, which always holds with $D = R + \|A\|_{\infty,2}$. Then B_R is invariant under F, and F has upper factor $q = b(1 + \kappa_f) + SD$. If $q < 1$, every initialization in B_R converges geometrically to the unique fixed point in B_R.*

Proof. Each node of $F(X)$ is a convex combination of points of norm at most R, so B_R is invariant. Node by node, $F(X)_i - F(Y)_i = a_i(X)\big((\mathcal{P}(X)X)_i - (\mathcal{P}(Y)Y)_i\big) + \big(a_i(X) - a_i(Y)\big)\big((\mathcal{P}(Y)Y)_i - A_i\big)$. The first term is at most $b(1+\kappa_f)\|X-Y\|_{\infty,2}$, as in Theorem 7.1; the second is at most $SD\|X-Y\|_{\infty,2}$. Banach's theorem finishes the proof. □

Constant shares, $S = 0$ and $b = \alpha$, give back the factor $\alpha(1 + \kappa_f)$ of Theorem 7.1. Letting the share adapt costs SD: how fast the share moves with the state, times how far the mixed state can sit from its anchor. The proposition is checked in Lean. It is a first step toward GRAND++ and deep equilibrium layers, whose other terms still need bounds of their own (Section 4.2).

Applying the certificate needs a bound on the builder. For softmax it is explicit.

Proposition 7.4 (Dot-product builders)**.** *For dot-product softmax with nonnegative logit scale s, $\mathcal{P}(X)_{ij} = \mathrm{softmax}_j(s\langle x_i, x_j\rangle)$, we have $L_P \le 2sR$ on B_R, so $\kappa_f \le 2sR^2$. Thus $2\alpha sR^2 < 1 - \alpha$ certifies convergence.*

The proof is in Section A.8. Stronger anchoring, softer logits, or a tighter invariant domain improve the test. With input initialization, all rows remain in the anchor convex hull C. Since builder differences have zero row sums, RL_P can be replaced by $D_C L_P/2$, where D_C is the hull diameter, and on the hull $L_P \le s(D_C/2+R)$ (Section A.8). Perturbed anchors still require a common domain. The unit ball certifies the dot-product setting $(\alpha, s) = (0.6, 0.3)$ with $m = 0.04$. It does not certify the strong-mixing image failure of Section 12.4 or the ordinary-kernel response maximum of Section 12.2. For each builder, twelve (α, s) pairs have nonzero mixing: the recorded radius $R = 1.25$ certifies the dot-product builder (K0) only at $(0.3, 0.3)$; $R = 1$, valid for the bounded prepared anchors and their prescribed zero or anchor starts, additionally certifies $(0.6, 0.3)$; and the stated global bound for the Gaussian builder with adaptive bandwidth (K2) is inactive at every nonzero pair. In the historical grid of prepared features and class probabilities (Section 12.5), context-specific hulls certify 1,349 of 4,608 nonidentity image confirmation configurations, all dot-product (Table 7.1). A hull certificate guarantees convergence on that context's invariant hull, including its anchor start;

Table 7.1: **Certified nonidentity configurations per phase.** The ball columns use the analytic bounds of Proposition 7.4 and Section A.8; the hull column uses the actual image anchors and $L_P \le s(D_C/2 + R)$ for K0, the tied dot-product builder. K2, the Gaussian builder whose bandwidth depends on the evolving cloud (Section 12.2), contributes zero throughout. Identity cases are excluded. The two image phases contain different sampled contexts.

Family	Phase	Total	$R = 1.25$	$R = 1$	Anchor hull
Geometry	Each	720	30	60	—
Images	Discovery	4,608	192	384	1,309
Images	Confirmation	4,608	192	384	1,349

it does not automatically certify starts outside the hull or give a full-neighborhood derivative bound. The book keeps certified convergence distinct from numerical settling throughout, and Chapter 13 develops the difference.

The certificate is checked by machine (Bruce, 2026h). A formal development in Lean 4 (version 4.33.1, with Mathlib) proves the invariant ball, the contraction factor and the unique fixed point of Theorem 7.1, with geometric convergence and uniqueness, for a stochastic builder with a common-ball Lipschitz bound; the anchor-response bound of Corollary 7.2, in the form $\beta/(1-q)$, for two anchors in the same invariant ball and the same builder; the dot-product bound of Proposition 7.4, for full unmasked dot-product softmax with the explicit ball radius; and the class-sharp obstruction, an explicit scalar contrast family with three fixed points and two local attractors arbitrarily near the threshold, with constructive C^1 regularity and the local attraction and repulsion of its fixed points. The formal sharpness theorem thus constructs a C^1 scalar family. The C^∞ mollification and the embedding into the token-state model are proved on paper (Section A.8), as are the hull refinement and the convex-hull invariance that follow Proposition 7.4. For the Gaussian builder with adaptive bandwidth, the formal development separates the representation feedback from the bandwidth feedback algebraically, and the numerical constant of the implemented builder is established on paper. An audit compiles all twelve modules and lists 56 declarations, foundational lemmas included, each using only the standard axioms (propositional extensionality, choice and quotient soundness), with no unfinished proofs and no added axioms; it counts declarations, not distinct theorems.

Past the margin, the witness of Theorem 7.1 has two attracting fixed points, and Section 7.6 returns to this loss of uniqueness. Gain control has a second removal signature: an input response that grows when the weights adapt, although the frozen average, a convex combination of the anchors, cannot amplify in the maximum-row norm (Section 1.4). It is the subject of Part III, and Chapter 12 states it once (Sections 12.1 and 12.2).

7.2 Softmax's feedback constant

For dot-product softmax on a radius-R ball the feedback constant is $\kappa_{\mathrm{f}} = RL_P \le 2sR^2$ (Proposition 7.4), so the anchor, the logit scale and the radius meet in one margin. With row mass at most $r_{\max}$, as in a head that can abstain (Chapter 8), the condition reads $q = \alpha(r_{\max} + RL_Q) < 1$ (Bruce, 2026d); with $r_{\max} = 1$ and $L_Q = L_P$, q is the factor $\alpha(1 + \kappa_{\mathrm{f}}) = 1 - m$ of Theorem 7.1.

Proposition 7.5 (Feedback gain with row mass)**.** *Use the maximum-row norm $\|X\|_{\infty,2} = \max_i \|x_i\|_2$ and the row-sum norm for matrices. On the closed ball B_R let $Q(X) \ge 0$ satisfy $\|Q(X)\|_\infty \le r_{\max}$ and $\|Q(X) - Q(Y)\|_\infty \le L_Q \|X - Y\|_{\infty,2}$. The update $T_A(X) = \alpha Q(X)X + \beta A$ is a contraction on B_R when $q = \alpha(r_{\max} + RL_Q) < 1$, and for $\alpha + \beta = 1$, $\|A\|_{\infty,2} \le R$ and $r_{\max} \le 1$ it maps*

the ball to itself. For sink-softmax weights with scores $s\langle x_i, x_j\rangle$ and a state-independent null logit, $L_Q \le 2|s|R\, r_{\max}$, so $q \le \alpha r_{\max}(1 + 2|s|R^2)$.

The proof is in Section A.8. If the null logit depends on the state, the relative scores $s_{ij} - b_i$ must be differentiated and a term in $\|\mathrm{D}b_i\|$ is added. An anchor that itself depends on the state with Lipschitz constant ξ lowers the margin $1 - q$ by $\beta\xi$. A row mass observed at one state does not certify the ball.

The constant in the bound matters for what it certifies.

Remark 7.6 (The loop-gain constant). For dot-product softmax, the constant 2 in $\kappa_{\mathrm{f}} \le 2sR^2$ is the tight one. The looser bound $8sR^2$, which the detector test of Section 7.4 used, is four times the tight one.

The proof is in Section A.8.

7.3 Weight decay as the anchor

If decoupled weight decay is modelled as the anchor, with the learning rate η as the step share ($\alpha = \eta$) and $\eta\lambda$ as the anchor share ($\beta = \eta\lambda$), the margin ties the logit scale to weight decay: tied dot-product softmax on a domain of radius R has loop gain at most $2sR^2$, the learning rate cancels, and the margin is positive below the cap $s_c = \lambda/(2R^2)$. The cap reads the AdamW step $W \mapsto (1 - \eta\lambda)W - \eta\nabla L$ as an anchored iteration. That is a model, not a case of Theorem 7.1: the theorem's two shares sum to one, while η and $\eta\lambda$ do not, so the cap rests wholly on the modelling assumption. Chapter 0 states the result (Section 0.4.3); its full treatment is left for later work.

What is already established about the threshold comes from exact, synthetic and in-system checks. In machine-precision checks on constructed loops, whose worst self-consistency residual is 8.9×10^{-12}, the fixed point below the threshold is unique and every step contracts by at most $q = \alpha(1 + \kappa_{\mathrm{f}})$; the loss of uniqueness is a saddle-node tangency at its closed-form location (worst tangency residual 1.1×10^{-9}); past the threshold three fixed points appear, the outer two attracting, with hysteresis as the anchor is swept; and on a builder that attains the bound, the relaxation time and the sensitivity to the anchor grow like $1/m$ as the margin closes (fitted exponents -0.99992 to -0.99996 and -0.9961 to -0.9998). In deterministic synthetic checks, rebuilding the weights from an exponential moving average of the states does not move the boundary. For a two-point kernel that attains the bound, two starts end apart exactly when $\alpha\kappa_{\mathrm{f}}/\beta$ is 1.05, 1.3 or 2 and agree when it is 0.5, 0.8 or 0.95, at averaging rates of 1, 0.3, 0.1 and 0.01 alike and in each of three seeds; the rate changes only the time to the outcome, which grows roughly as its inverse (fitted exponents 0.83 and 0.85 at the ratios 0.8 and 1.3). A softmax Gram kernel calibrated to the same target ratios brings the two tested starts together within the test's tolerance at every one of them. A smoother whose bandwidth is rebuilt from the moving states runs away, its bandwidth falling by a factor of 1.85×10^{17} within twenty steps, while the same smoother with normalized states and a fixed bandwidth keeps its scale at one. The in-system checks are partial. In a trained rig, with the loop gain measured on frozen embeddings, the measured contraction factor stayed within $\alpha(1+\kappa_{\mathrm{f}})$ in every epoch of the 14 cells with a positive median margin, and the sustained loss-increase criterion was met in none of the 72 cells of its sweeps, those with negative margins included. In a teacher–student loop with the anchoring channel closed and the loop gain set by sharpening the teacher, all nine cells that learned had a negative minimum margin and collapsed, while the other fifteen did not learn. No cell both learned and kept a positive margin, so only the side past the threshold is confirmed, by a re-scoring, motivated after the fact, of runs whose first reading was inconclusive.

7.4 Divergence flags do not test the coupling

A divergence detector can flag the anchor's relaxation as feedback. In a synthetic loop test of this coupling, the boundary and margin sweeps use five seeds. A relaxation of the anchor without any feedback reproduces all 170 of the detector's flags, with one false alarm when predicting which of the 550 cells were originally flagged, so the flags measure the relaxation. Chapter 0 states the result (Section 0.4.3), and its full record is left for later work: the threshold is proved, and the cap it implies is untested in language models.

7.5 Norm control manages gain without certifying it

The anchor, the logit scale and the representation radius trade against one another in the margin. The statement concerns the full feedback Jacobian: normalization bounds the radius but is not by itself a certificate. Among the changes to the mixing layer and its block that practice has proposed or adopted, four read as gain control: QK-norm (2020), z-loss (2022), σReparam (2023) and logit soft-capping (2024), three of them bounding s in $\kappa_{\mathrm{f}} \le 2sR^2$ and z-loss acting on the output logits, a block-level reading (Henry et al., 2020; Chowdhery et al., 2023; Zhai et al., 2023; Gemma Team, 2024). The years are those of first public release, not of adoption, and the reading is an interpretation, not evidence; it is one row of the retrodiction that Section 16.5 prints.
Pending: Coupling the logit scale to the anchor, enforced by a spectral cap on the query–key map, against the standard fixes at 125M parameters: maximum stable learning rate, learning-rate sensitivity, instability incidence and final loss without a fix, with the coupling, and with QK-norm, logit soft-capping, z-loss, σReparam, and QK-norm with z-loss.

7.6 Gain control is uniqueness

When routing is rebuilt from the machine's own output, the loop has gain. If the gain beats the tether, fixed points can multiply, and the output can then depend on where the iteration started. Removing gain control therefore shows up as a loss of uniqueness (Bruce, 2026e). Consider the tethered attention map $F(X) = \beta X_0 + \alpha P(X)X$, with $\beta = 1 - \alpha$ and $P(X)$ the row softmax of $\tau\, x_i^{\top} x_j$ over fixed visible sets V_i of sizes m_i; self may be excluded, but the pattern does not change with X. It is the iteration (1.4) with the input X_0 as its anchor, the tether of Section 5.4, and a dot-product builder with logit scale τ and a fixed mask. Write $\|X\|_{\infty,2} = \max_i \|x_i\|_2$ for the row radius of X, $\max_{i,j} \|x_i - x_j\|_2$ for its row diameter, $R_0 = \|X_0\|_{\infty,2}$, and x_{0i} for the rows of X_0.

Proposition 7.7 (Gain control)**.**

(a) The basic certificate. *The ball B_R with $R \ge R_0$ is invariant, and on it* $\mathrm{Lip}(F) \le \alpha(1 + 2\tau R^2)$. *If $\alpha(1 + 2\tau R^2) < 1$, the ball holds a unique fixed point, which attracts every start in it.*

(b) A smaller absorbing set. *Take any certified lower bound $0 < \ell_i \le e^{-2\tau R_0^2}/m_i$ on the routing probabilities, and set $c_i = \beta x_{0i} + \alpha\beta\ell_i \sum_{j \in V_i} x_{0j}$ and $r_i = \alpha(1 - \beta m_i \ell_i) R_0$. The product of balls $\mathcal{K} = \prod_i \overline{B}(c_i, r_i)$ is invariant, lies inside the ball B_{R_0}, and contains every trajectory from that ball after two steps. So $\alpha(1 + 2\tau R_{\mathrm{eff}}^2) < 1$, with $R_{\mathrm{eff}} = \max_i(\|c_i\|_2 + r_i) \le R_0$, certifies a unique fixed point in the original ball, attracting every start in it.*

(c) A local certificate. *For a candidate $\widehat{X}$ with residual $e = \|F(\widehat{X}) - \widehat{X}\|_{\infty,2}$, a radius $r > 0$, $R = \|\widehat{X}\|_{\infty,2} + r$ and $q = \alpha(1 + 2\tau R^2)$: if $q < 1$ and $e \le (1 - q)r$, the ball of radius r around $\widehat{X}$ holds a unique fixed point, which attracts every start in that ball and lies within $e/(1 - q)$ of*

$\widehat{X}$. On a convex domain of row radius R and row diameter at most D, q may be replaced by $\alpha(1+\tau RD)$.

The proof is in Section A.8.

The certificates are sufficient, not necessary. A radius measured at a fixed point certifies nothing by itself. Part (b) of Proposition 7.7 builds an invariant set before any radius is substituted, and part (c) certifies only its own ball, and only with a verified residual. Nor can a smaller ball centred at zero be invariant in general: if every row is ru, with u the direction of an input row of norm R_0, that input's output row has norm $\beta R_0 + \alpha r > r$. For unit-norm inputs the basic certificate holds for $\tau < (1-\alpha)/(2\alpha)$, a threshold that vanishes as $\alpha \to 1$. If the true onset of multistability stays bounded away from zero, the certificate's conservatism must therefore grow at least like $\alpha/(1-\alpha)$. That is a conditional explanation of a growing gap, not a law for where the transition sits, and we leave its location open. The gain-control condition of Sections 7.1 and 7.2 is the full-map form of this sufficient condition (Bruce, 2026d).

Constructed-key records. A gain-control test iterates the map from two starting states on 64 unit-norm tokens in 16 dimensions, with self excluded and three seeds (Figure 5.2b; Bruce, 2026e). The two starts end within a relative Frobenius distance of 10^{-6} of each other, in every seed, up to $\tau = 11.3$, 8.0, 16.0 and 22.6 at $\alpha = 0.3$, 0.5, 0.7 and 0.9. There the basic certificate stops at $(1-\alpha)/(2\alpha) = 1.17$, 0.50, 0.21 and 0.06. It is conservative by a factor of about 9.6 at $\alpha = 0.3$, growing to about 400 at $\alpha = 0.9$ (16 and 74 between). The test's τ grid steps by factors of $\sqrt{2}$, so each factor, rounded down to two figures, is a lower bound within that step. These readings are provisional. The test ran below its registered size (128 tokens in 32 dimensions). 7 of its 252 cells did not converge within 1,500 iterations, two of them with an estimated local gain above one (1.08 and 1.07). And a second start's convergence was not recorded, so the readings show endpoint dependence, not a certified comparison of fixed points. The accepted recovery now includes the registered $N = 128, d = 32$ test with five seeds, a 5,000-update cap, the 1,500-step state and both starts' convergence flags. Its saved median first-loss values are $\tau_{\text{loss}} = 16.000, 16.000, 22.627, 45.255$ at $\alpha = 0.3, 0.5, 0.7, 0.9$; the registered check on this band passes: at every α the median lies within a factor of two of the historical median τ_{loss} (16, 11.31, 22.63 and 32). Figure 5.2b shows every seed at both sizes. The source paper's gain-certificate table retains all per-seed readouts and saved medians/minima/maxima (Bruce, 2026e, Table 5 and Appendix C.1). It distinguishes the dimensionless ratios $\tau_{\text{last}}/\tau_{\text{cert,reg}}$ and $\tau_{\text{last}}/\tau_{\text{cert}}$ from the Lipschitz value $\alpha(1+2\tau_{\text{last}}R^2)$, at both domain radius $R = 1$ and realized radius $R = R_{\text{real}}$. Here τ_{last} is the preceding unique grid point, $\tau_{\text{cert}} = (1-\alpha)/(2\alpha R^2)$ and the registered threshold has denominator $4\alpha R^2$. The extrema are seed ranges, not confidence intervals. A realized radius alone supplies no invariant set; endpoint separation is not a certified comparison of fixed points. The saved execution failure remains, and publication is admitted as a rebaseline, not historical reproduction.

7.7 Mechanism and onset in trained language models

Pending: The mechanism, tested by an experiment that separates feedback from the anchor's relaxation: where instability sets in, in the plane of logit scale and weight decay, for a softmax Gram loop, against the certified boundaries ($2sR^2$ and $8sR^2$) and the boundary set by the measured gain, with a frozen-kernel control and a control that attains the bound.

Pending: Where logit growth destabilizes language-model training: instability incidence against learning rate and weight decay at proxy scale and at 125M parameters, with the contour that the certificate predicts.

7.8 What holds in trained networks

Exact under premises checked network by network. Gain control (Sections 7.1, 7.2 and 7.6): the per-layer bound $\kappa_{\mathrm{f}} \leq 2sR^2$ holds for every softmax layer with tied dot-product scores (projections add their norms). Contraction and a single attracting state are statements about one builder applied repeatedly, exact once $m > 0$ is computed on it: weight-tied loops and deep equilibrium models are in that class, a stack of distinct layers applied once each is not. The certificate for adaptive averaging (Section 7.1): exact for any anchored adaptive iteration once $m > 0$ is computed on it: weight-tied loops, deep equilibrium models, iterative denoisers and test-time mixing. A stack of distinct layers applied once each is not in the class. The logit-scale cap under weight decay (Section 7.3): the loop-gain bound holds for every softmax layer with tied dot-product scores (projections add their norms); the cap assumes that decoupled weight decay acts as the anchor ($\alpha = \eta$, $\beta = \eta\lambda$), a modelling assumption on which it rests wholly, since these shares do not sum to one as Theorem 7.1 requires. Chapter 0 explains why the certificate is the clearest case of the difference between one operator applied over and over and a different operator at each layer (Section 0.5).

What a converged average guarantees depends on the quantity that is used. Table 7.2 separates four questions, their mathematical objects, and the evidence each requires (Bruce, 2026h). Sharpness holds over the stated builder class; kernel-specific thresholds remain open, and global certificates can be conservative. The RED/NLM probe establishes possibility (Section 12.3). The classification study isolates cross-image mixing with one pretrained encoder, dataset and fixed heads (Section 12.4). The prevalence and magnitude of this interference in trained architectures remain open. Learned projections, multi-head structure, normalization and patch-level mixing require their own analysis. The complementary controls on 16-dimensional features have incomplete historical extraction provenance; the documented full-dimensional study has a separate scope. Positive weights constrain output values; rebuilding introduces another feedback path. Table 7.2 gives a practical sequence: compute m on a common domain where builder bounds permit; compare adaptive and frozen responses and class margins; examine H_K at the iteration budget actually consumed. Convergence, input response, task fidelity and downstream stability each require their own evidence.

Table 7.2: **Match the guarantee to the quantity being used.** The bounds use the maximum-row norm and a common invariant domain; $\|Dg\| \leq 1$ is required for the all-budget result, where $X_0 = g(A)$ is the inner initialization of Section 13.5. $H = D_A X^*$ is the response of the equilibrium to its anchors (Section 12.1), with DF the update's Jacobian; H_K is the derivative of K inner updates, and γ the strength of an outer feedback loop that uses them (Section 13.5).

Question	Relevant object	Supported conclusion
Does the solve converge?	$m = \beta - \alpha R L_P$	$m > 0$ guarantees a unique limit; the class threshold is sharp.
Does the limit amplify input changes?	$H = \beta(I - DF)^{-1}$	Gain can approach β/m even as the positive frozen map R_* approaches identity.
Are predictions preserved?	Own and companion margins	The margin condition of Section 12.5 certifies class preservation; context removal tests recoverability.
Is downstream feedback stable?	The finite derivative H_K	Intermediate budgets can expand; $\gamma < m/\beta$ suffices for every budget.

Chapter 8

A way to deliver nothing: abstention

Chapter 0 names abstention as the fourth condition (Section 0.3.2): softmax weights add up to one, so a head must deliver something even when nothing is relevant. Section 1.1 set aside the kernels whose rows may sum to less than one, and this chapter takes them up. Section 8.1 locates the constraint in the softmax Jacobian, and Section 8.2 shows that the zero-valued null patches and headwise output gates that let a head attend less are one operation, a gate times an ordinary softmax, which reads as a killed Markov chain. Section 8.3 reads attention rollout as survival and sinks as skip switches. Section 8.4 shows that whether a null keeps abstention at every length is decided by the test it applies, not by the machine that applies it, and Section 8.5 reads trained models against the constraint. Section 8.6 records a prediction for test-time mixing, and Section 8.7 says what holds in trained networks. What a null that grows with the context costs retrieval is the two-sided law of Chapter 9.

8.1 Sum-to-one leaves no direction for attending less

Fix one query and write its N logits as $s \in \mathbb{R}^N$. A softmax head assigns $p = \mathrm{softmax}(s)$, $p_j = e^{s_j} / \sum_k e^{s_k}$. Its Jacobian is

$$\frac{\partial p}{\partial s} = \mathrm{diag}(p) - pp^\top, \tag{8.1}$$

which is positive semidefinite, and for positive weights its kernel is exactly the span of $\mathbf{1}$. The kernel is the uniform shift $s \mapsto s - c\mathbf{1}$, the one change of the logits that could mean "attend less", and it changes nothing. A softmax head cannot lower its total attention, and gradient descent cannot teach it to: the direction carries no gradient. A null logit b changes this. The head becomes

$$p_j = \frac{e^{s_j}}{e^b + \sum_k e^{s_k}}, \qquad r = \sum_j p_j < 1, \tag{8.2}$$

where the missing mass $1 - r$ is the head's abstention. Off-by-one softmax is $b = 0$ (Miller, 2023); a learned sink is a trained b per head (OpenAI, 2025). Several value-null registers, whose values are zero, enter through their summed exponential weight. The Jacobian keeps the form (8.1), now with $\mathbf{1}^\top p = r$. Write $\mathrm{LSE}(s) = \log \sum_k e^{s_k}$ for the log-sum-exp and $\sigma(z) = 1/(1 + e^{-z})$ for the logistic function.

Proposition 8.1 (A null makes the Jacobian definite)**.** *For the head* (8.2) *and every* $x \in \mathbb{R}^N$,

$$x^\top \frac{\partial p}{\partial s} x \;\geq\; (1 - r) \sum_j p_j x_j^2,$$

so the Jacobian is positive definite with $\lambda_{\min}(\partial p/\partial s) \geq (1-r)\min_j p_j$. *It is the Hessian of the convex potential* $s \mapsto \log(e^b + \sum_j e^{s_j})$. *Moreover* $p = r\,\mathrm{softmax}(s)$ *with* $r = \sigma(\mathrm{LSE}(s) - b)$.

The proof is in Section A.9. The inequality is Cauchy–Schwarz, $(\sum_j p_j x_j)^2 \leq r\sum_j p_j x_j^2$; the other statements are direct computations. So abstention and the convexity of log-sum-exp coexist: a head can learn to attend less, and its weights are still the gradient of a convex potential. For a numerical instance with $r = 0.953$, the smallest eigenvalue is 3.8×10^{-3}, above the bound $(1-r)\min_j p_j = 3.1 \times 10^{-4}$; without the null it is zero to machine precision. The gain that a feedback loop through such a head can have scales with its row mass r as well. Companion work studies that sensitivity for heads whose rows sum to one (Bruce, 2026h). There is also a norm picture. A softmax head over a sequence is a row-stochastic matrix P with a stationary law π, which is positive when the head is unmasked. Measure vectors by $\|v\|_\pi$, the square root of the π-weighted average of their squared components. Jensen's inequality with stationarity gives $\|Pv\|_\pi \leq \|v\|_\pi$ with equality on constants, so $\|P\|_\pi = 1$ exactly (Lemma 1.2(ii)): the head can never shrink the constant direction. An abstaining head is a substochastic matrix Q, and when it is irreducible with at least one deficient row it is a strict contraction, $\|Q\|_w = \rho < 1$ in a weighted norm built from its Perron vectors (Theorem 8.3). Stochasticity forbids one more thing: a stochastic head cannot output zero for every collection of identical nonzero values. It can output zero by cancellation between values or through an output map, which is a different mechanism.

8.2 Every null patch is a gate times a softmax

The mechanisms that let a head abstain (Bruce, 2026c, Section 2) look different, but most of them are the same operation. Write $\mathrm{LSE}_i(s) = \log\sum_j e^{s_{ij}}$ for row i of a logit matrix s.

Proposition 8.2 (Gate times routing). *A head with a null weight* $z_i > 0$ *in row* i, $P_{ij} = e^{s_{ij}}/(z_i + \sum_k e^{s_{ik}})$, *factors as*

$$P = \mathrm{diag}(g)\,\mathrm{softmax}_{\mathrm{row}}(s), \qquad g_i = \sigma(\mathrm{LSE}_i(s) - b_i), \quad b_i = \log z_i.$$

Off-by-one softmax is $b_i = 0$; *a learned sink is* $b_i = b$, *one trained number per head; value-null registers give a query-dependent* b_i, *the log of their summed exponential weight in row* i.

The proof is in Section A.9. A headwise output gate (Qiu et al., 2025), $o_i \leftarrow g_i o_i$ with $g_i = \sigma(w^\top x_i + c)$, has the same form with the gate driven by the query's input instead of its logits. So every sink-logit or headwise-gated head routes by an ordinary softmax and then scales each row by one number in $(0, 1)$: the gate decides how much the head says, the softmax decides where it looks. What differs between the mechanisms is only what drives the gate. Concurrent work (Wang, 2026) decomposes a gate on the values the same way, into a renormalized routing times a row attenuation (their Lemma 3.1). Sigmoid attention (Ramapuram et al., 2025), $P_{ij} = \sigma(s_{ij} + c)$, is outside the class: each key has its own gate, the row mass is not bounded by one, and nothing couples the keys. Its behaviour under the log-N law of Chapter 9 is therefore different, and Proposition 9.3 treats it separately.

An abstaining head is a killed Markov chain. Row i keeps mass $r_i = g_i$ and loses $1 - r_i$, so the head is a substochastic kernel Q. Conditioning on survival turns it back into attention. Here $D_x = \mathrm{diag}(x)$, and $x \circ y$ is the entrywise product.

Theorem 8.3 (Killing, conditioning and reversal as attention). *Let* $Q \geq 0$ *be nonzero, irreducible and substochastic with some row sum below one, with positive right and left Perron vectors* h *and* ν *and Perron root* ρ. *Then* $0 < \rho < 1$, *the conditioned kernel* $P^h = \rho^{-1}D_h^{-1}QD_h$ *is stochastic*

with stationary law $\pi^h \propto \nu \circ h$*, its reversal is* $(P^h)^\dagger = \rho^{-1} D_\nu^{-1} Q^\top D_\nu$*, and* $\|Q\|_w = \rho$ *in the norm* $\|x\|_w = \|D_h^{-1} x\|_{L^2(\pi^h)}$*. If* $Q_{ij} = e^{s_{ij} - L_i}$*, then*

$$P^h = \mathrm{softmax}_{\mathrm{row}}(s + \mathbf{1}(\log h)^\top), \qquad (P^h)^\dagger = \mathrm{softmax}_{\mathrm{row}}(s^\top + \mathbf{1}(\log \nu - L)^\top).$$

The proof is in Section A.9. The conditioned kernel P^h, the Doob transform of Q, is an ordinary softmax head with a key bias $\log h$, and its reversal is again an attention call. A path conditioned on survival to a finite time generally has a different law from a path of this kernel, as the proof of Proposition 8.4 shows. The Perron pair is computable with log-sum-exp alone, since $\log(Qh)_i = \mathrm{LSE}_i(s + \mathbf{1}(\log h)^\top) - L_i$; in a numerical check on a 12-token off-by-one head, the corresponding iteration on the transposed logits gives ν and ρ to machine precision in 39 steps. For a primitive Q, the endpoint conditioned on survival converges to $\nu/\nu^\top \mathbf{1}$, while the path occupation, the fraction of time a surviving path spends at each token, converges to π^h, in general a different law.

Proposition 8.4 (Endpoint and occupation limits)**.** *If* Q *is also primitive and* $\nu^\top h = 1$*, then for any initial law* μ*, conditioned on survival to time* t*, the endpoint law converges to* $\nu/\nu^\top \mathbf{1}$*, and the empirical path occupation converges in probability to* $\pi^h = \nu \circ h$*.*

The proof is in Section A.9. The proofs use standard finite-state Perron and Markov-chain facts; we make no novelty claim for them, and the reversal machinery follows Fill (1991) and Section 3.1.

The conditioned head is an ordinary attention head, so it has the three parts of Section 3.2. Its stationary law π^h is strictly positive, so P^h splits orthogonally into preference, exchange and circulation. For finite scores, its scores differ from s only by the key bias $\log h$, a potential, so they carry exactly the cycles of s (Theorem 3.2): the conditioned head circulates if and only if the original scores carry a cycle. This is checked in Lean. It needs the positive Perron pair of Theorem 8.3, which irreducibility supplies; a causal head's survivor kernel is triangular and in general has none, so decoders keep the reading of Section 2.5.

Decoders are killed by construction. A causal mask supplies an absorbing state without any null. In a causal softmax head the first token attends only to itself, so e_1 is stationary and the spectrum is the diagonal. Tracking only the other tokens gives a killed chain: mass sent to the first token is lost from that chain. Every other row gives token 1 positive mass, so every transient token reaches absorption, the transient block Q has $\|Q^m\|_\infty < 1$ for some m, and repeated application drives all mass to token 1 (Section 2.5). The size of this structural sink is large in practice. Its census, over seven public causal language models, is in Section 2.7; there the minimum row mass sent to token 1 falls below 10^{-6} in only a small share of the operators, all of them heads, and the smallest of these values are stored as exact zeros in float16. One caution: conditioned on survival, a causal chain concentrates on the transient token with the largest self-attention only if that token is reachable from where the chain starts. The chain $Q = \mathrm{diag}(0.9, 0.5)$ started in state 2 stays there whenever it survives.

8.3 Rollout measures survival, and sinks act as skip switches

Causal attention changes the object of study. Causal kernels are lower-triangular, which makes the first token absorbing (it attends only to itself); they are also far from normal, and each layer has its own. There is no common spectrum, but there is a common target, and the natural quantity becomes survival. Let $A_1, \ldots, A_L$ be the head-averaged attention kernels of a causal stack on tokens $1, \ldots, T$, and let $p_\ell \in [0,1]^T$ hold the share of the attention branch in each token's update. Attention

rollout (Abnar & Zuidema, 2020) composes the layers

$$R_\ell = \operatorname{diag}(\mathbf{1} - p_\ell) + \operatorname{diag}(p_\ell)\, A_\ell, \qquad T_L = R_L \cdots R_1, \tag{8.3}$$

with the convention $p_\ell \equiv \frac{1}{2}$. Residual-aware attribution methods instead use network-dependent contributions (Kobayashi et al., 2021; Ferrando et al., 2022; Modarressi et al., 2022; Mohebbi et al., 2023). Row i of T_L is the law of a walk that starts at output token i and, at layer $\ell = L, \dots, 1$, stays put with probability $1 - p_{\ell,j}$ or jumps according to A_ℓ. Write Q_ℓ and $B_\ell = \operatorname{diag}(\mathbf{1} - p_\ell) + \operatorname{diag}(p_\ell) Q_\ell$ for the blocks of A_ℓ and R_ℓ on tokens $2, \dots, T$, and $s_{\ell,i} = (A_\ell)_{i1}$ for the attention that token i pays to the first token.

Theorem 8.5 (Rollout is survival)**.** *The distance of rollout from first-token collapse is exactly twice the worst-case probability that the walk avoids the first token:*

$$\tfrac{1}{2}\left\|T_L - \mathbf{1}e_1^\top\right\|_\infty \;=\; \left\|B_L \cdots B_1 \mathbf{1}\right\|_\infty \;=\; \max_i \mathbb{P}_i(\textit{no absorption in } L \textit{ layers}). \tag{8.4}$$

The survival vector obeys $\prod_\ell (1 - \max_i p_{\ell,i}) \le \|B_L \cdots B_1 \mathbf{1}\|_\infty \le \prod_\ell (1 - \min_i p_{\ell,i} s_{\ell,i})$*, and the upper bound is attained by self-or-sink kernels with shares constant across tokens. For a single kernel* A *and an independent proper path law,* $\|f(A) - \mathbf{1}e_1^\top\|_\infty = 2 \max_i \mathbb{E}[(Q^K \mathbf{1})_i]$*, and a sink leak* $s_i \ge \delta$ *for all* i *gives the sharp envelope* $2\mathbb{E}[(1-\delta)^K]$*.*

Proof. Each row of the stochastic matrix T_L has ℓ_1 distance $2(1 - (T_L)_{i1})$ from e_1, and the mass off the first token is the row sum of the transient block $B_L \cdots B_1$. The envelopes follow from $(1 - p_{\ell,i}) \le (B_\ell \mathbf{1})_i = 1 - p_{\ell,i} s_{\ell,i}$ and nonnegativity. The full proof is in Section A.9. □

The identity needs no assumption on queries, keys or values: it holds for the realized kernels of any causal stack, per input and per head. Its single-kernel clause is the identity and envelope of Theorem 2.16, and with one share per layer its envelopes are the bounds of Proposition 2.20(i). In survival-analysis terms, $p_{\ell,i} s_{\ell,i}$ is the hazard of token i at layer ℓ, and the envelopes show that the residual share sets the time scale together with the attention pattern. On seven batch-averaged, 64-token attention captures, lowering a common share from $p = \frac{1}{2}$ to $p = 0.1$ delays the first depth at which worst-token survival reaches $\frac{1}{2}$ by a factor of 2.1–3.2 in the five checkpoints that reach it, and beyond the model's depth in the other two (Section 2.7). Wu et al. (2025) prove that causal masking concentrates deep attention on early positions under assumptions on query, key and value products, and Karagodin et al. (2024) prove a continuous-time analogue with identity value maps, in which every token converges to the first for almost every initial configuration; Theorem 8.5 is an identity that makes the concentration measurable. Designs whose attention rows sum to less than one are covered by adding one virtual absorbing token that receives the missing mass. Examples are a learned sink logit in the softmax denominator (OpenAI, 2025), the off-by-one softmax (Miller, 2023), stick-breaking attention (Tan et al., 2025). Unnormalized sigmoid attention (Ramapuram et al., 2025) is covered only when every row has total mass at most one.

Trained language models place much of their attention on the first token (Xiao et al., 2024; Gu et al., 2025; Barbero et al., 2025a), and this mass often carries little value (Guo et al., 2024). Concentration of rollout on the sink is therefore a statement about where provenance flows; by itself it says nothing about lost information. What a system does with the sink changes the diagnostic, and three treatments are in use. Evicting the sink from the KV cache renormalizes each softmax over the remaining tokens; this is what destroys the perplexity of window attention (Xiao et al., 2024). A zero-value sink keeps the attention mass but writes nothing, as with gpt-oss's learned sinks (OpenAI, 2025), the off-by-one softmax (Miller, 2023) and the key-bias sinks of Gu et al. (2025). Treating attention to the sink as a skip returns that mass to the token itself. For one layer,

with s the sink column, $s_i < 1$ in every content row, and Q the content block, these give three stochastic content diagnostics. For a zero-value sink the killed rollout operator on content is B; C below renormalizes each surviving row at every layer:

$$\underbrace{X = \operatorname{diag}(\mathbf{1} - p) + \operatorname{diag}(p)\operatorname{diag}(Q\mathbf{1})^{-1}Q}_{\text{eviction}}, \qquad \underbrace{C = \operatorname{diag}(B\mathbf{1})^{-1}B}_{\text{zero-value sink}}, \qquad \underbrace{F = B + \operatorname{diag}(p \circ s)}_{\text{sink as skip}}. \tag{8.5}$$

Proposition 8.6 (Sink treatments are ordered)**.** *For causal stacks whose content rows satisfy $s_{\ell,i} < 1$, with any per-token shares, every content token keeps weight on itself in the order*

$$(F_L \cdots F_1)_{ii} \;\geq\; (C_L \cdots C_1)_{ii} \;\geq\; (X_L \cdots X_1)_{ii},$$

with layerwise differences $F_{ii} - C_{ii} = p^2 s(1 - s - d)/(1 - ps)$ and $C_{ii} - X_{ii} = p(1 - p)s(1 - s - d)/[(1 - ps)(1 - s)]$, where $s = s_i$, $d = Q_{ii}$ and $p = p_i$. Under a zero-value sink the probability that token i's mass leaves it at layer ℓ is

$$\pi_{\ell,i} = \frac{p_{\ell,i}(1 - s_{\ell,i} - d_{\ell,i})}{1 - p_{\ell,i}s_{\ell,i}}, \tag{8.6}$$

which is nonincreasing as the sink mass $s_{\ell,i}$ grows with $p_{\ell,i}$ and $d_{\ell,i}$ held fixed, and $(C_L \cdots C_1)_{ii} = \prod_\ell (1 - \pi_{\ell,i})$.

The proof is in Section A.9. Causality makes the diagonal of a product the product of the diagonals, since a walk that moves to an earlier token cannot return, and the layerwise differences are then direct algebra. Two readings follow. First, eviction removes the brake. The attention that the sink absorbed is redistributed over content tokens, and every token keeps the least weight on itself. This is consistent with the collapse of window attention when initial tokens are evicted (Xiao et al., 2024), and the ordering follows from the proposition for every causal stack, whatever its attention. Second, (8.6) says that a zero-value sink is a *learned skip switch*: at fixed residual share and self-attention, increasing sink mass cannot raise the per-layer mixing hazard, and its self-weight survives depth if and only if $\sum_\ell \pi_{\ell,i} < \infty$ and no layer sends it away completely ($\pi_{\ell,i} < 1$ for every ℓ). Attention sinks are how a trained network can control its mixing budget (Section 2.4), which quantifies the argument that sinks prevent over-mixing (Barbero et al., 2025a); Section 2.7 measures how much they do.

8.4 The null's test decides abstention

A router that can retrieve must also be able to say "none of these". Whether it can do so at every length depends on how its null decision is made, not on what kind of machine makes it. Below, τ is the logit scale, a matching key scores c, and $m(\tau) = \mathbb{E}e^{\tau z}$ is the mean weight of a background score z, as in the mean-field model of Section 9.3.

Proposition 8.7 (The test decides)**.**

(a) A competitive null. *A null of weight e^b that competes with the background sum $B = \sum_j e^{\tau s_j}$ has abstention mass $a = e^b/(e^b + B)$, and $a \geq \frac{1}{2}$ exactly when $e^b \geq B$. In mean field, $B = Nm(\tau)$, so it fails at $\ln N = b - \ln m(\tau)$. The schedule $b = b_0 + \ln N$ keeps the predicted a at $e^{b_0}/(e^{b_0} + m(\tau))$ for every N. But a matching key's weight is then $e^{\tau c}/(e^{\tau c} + N(e^{b_0} + m(\tau)))$, which vanishes, and retrieval fails at $\ln N = \tau c - \ln(e^{b_0} + m(\tau))$.*

(b) An absolute test. *A rule that abstains when the largest score is below a threshold θ abstains with probability $F(\theta)^N$ for independent background scores with continuous distribution function F. This is one half at $\ln N = \ln\ln 2 - \ln(-\ln F(\theta))$. If every background score lies below θ almost surely, it abstains at every N.*

(c) A sampled test. *A router that checks only M sampled items, given the row, finds no item above θ with probability $(1 - p_\theta)^M$, where p_θ is its routing mass above θ.*

The proof is in Section A.9. The difference is what the decision sums over. Part (a) is the law of sums: a null that enters a normalization over all N items must grow like $\ln N$. A null that grows that way then costs retrieval, which is the two-sided law of Sections 9.1 and 9.2 (Bruce, 2026c,d). Two recent designs let the null grow with length: one tempers every score, its null included, by a learned factor and raises the null like the background's extreme value (Akbar, 2026), and the other trains a sink calibrated to the corpus size (Gollapudi et al., 2026). Part (b) reads a maximum. Its failure is set by the tail of the background at θ, not by a count, so it needs no schedule. It still fails where the background maximum reaches θ, which is the evidence wall of Section 9.3 in another form. Both parts read the background as exchangeable: (a)'s mean field takes B to grow like N, and (b) takes the background scores to be independent draws from a distribution that stays fixed as N grows. Positions break both in trained models. Inside the training length a rotary model's background sum grows like N^γ, with γ the measured growth, well below one for a fixed null, so by (a) a null that keeps abstention level needs only $\gamma \ln N$ (Section 9.2). Beyond it, positions that are not interpolated raise the background maximum with length, and a maximum then fails as that maximum grows (Section 17.4). A finite number of competitive walkers (routers that sample from the routing distribution, Section 9.3) realizes (a)'s decision as a binomial absorption event, which approaches the half-mass decision as the samples grow. A fixed number of verifying walkers, by (c), is not an exhaustive maximum: its abstention depends on its routing mass above θ, so it holds only while that mass stays small. In the constructed tests it held to 2^{30} items, though with weak evidence ($c = 0.5$) at $\tau = 30$ it was still falling there. On a present query its routing mass on the target falls with N, so it inherits the gain wall for retrieval.

Constructed-key records. The abstention test of Bruce (2026e) calibrates each null at $N = 64$ on the product substrate of Section 9.3 (Figure 8.1).

- *Competitive nulls.* The fixed softmax sink loses abstention within 0.43 nats of (a)'s onset in all 4 configurations. The walker's competitive killing loses it within 0.11 nats of the sink: the machines differ and the test is the same.
- *The scheduled null.* The sink $+ \ln N$, run for softmax only, keeps abstention to 2^{30} items with exact queries, and loses retrieval instead: at $\ln N = 6.24$ and 10.40 for $\tau = 10$ and 20, against 6.17 and 10.52 from (a) with this substrate's background sum in place of $Nm(\tau)$, as for the gain wall in Section 9.3; the comparison values were derived after these records were read.
- *Absolute tests.* Verifying walkers and the tree's key check hold abstention to 2^{30} items with exact queries. With 64 walkers, retrieval fails at $\ln N = 13.94$ ($\tau = 10$), as (c) implies, while with $M = N$ walkers it does not.
- *Weak evidence ($c = 0.5$).* The exhaustive absolute test, the tree's key check, loses abstention first, at $\ln N = 14.68$ and 14.63 ($\tau = 20$ and 30), before retrieval at 17.98 and 18.08. The largest background cosine, averaged over queries, reaches the threshold near $\ln N = 14.57$ on the test's grid, which is where a test on the maximum fails. Product keys share their halves, so their scores are not independent and (b)'s formula does not apply as written. The tempered sink, which keeps both sides to 2^{30} items with exact queries, loses abstention at 9.77 and 13.58: tempering needs its margins (Section 9.2; Bruce, 2026c).

A thought exercise behind these tests claimed that only address-based routers keep abstention

calibrated without a separate null schedule. The tests' registration predicted that the claim fails as stated, and the tests confirm it: verifying walkers, a sampling machine, keep abstention too. What keeps abstention without a schedule, up to its own evidence wall, is a test on a per-item statistic, whatever machine runs it. The accepted recovery also includes competitive killing with a $\ln N$ schedule: it preserves abstention and loses retrieval at the product mean-field onset. The four saved retrieval crossings, in the figure's row order, are 6.135, 10.488, 5.793 and 6.835 nats, against registered predictions 6.133, 10.488, 5.897 and 6.924. The saved competitive-null and scheduled-walker checks pass. These are recovered readouts under the registered publication exception; the overall assessment and execution gate remain failed, and the historical package was not reproduced (Bruce, 2026e, Appendix C.1).

8.5 What trained models already show

Small models have been trained that change one thing about how a head attends, and the sinks, the massive activations and the quality that result have been measured (Table 8.1; Bruce, 2026c). We read them against Section 8.1: a head given a way to attend less should need its sinks less. The models are small and the seeds few. Where a comparison has five seeds we give the paired change with its 95% interval; with two or three seeds we give the means and the number of seeds that moved in the stated direction, and every seed's value is in Table 8.2. We give no p-values.

The readouts. Massive activation is the largest hidden-state magnitude divided by the median one, per layer, then the maximum over layers (for sigmoid attention and the registers) or the mean (for the output gate and the value-null register). In the bidirectional ViTs a sink is a token with high stationary mass and a small value norm; the detector's third test, a high loading on the constant mode, ranks tokens exactly as stationary mass does there. In the causal models the stationary test is vacuous: the first token absorbs the chain, every other token's stationary mass sits at the numerical floor, and every token passes. The loading test is numerically degenerate there as well, yet its top-decile cut still selects. So a causal model's content-sink mass rests on the value-norm test and on that degenerate loading: read it as attention to a detector-selected subset of tokens with small value norms, not as mass on stationary-law sinks. First-token mass, where recorded, is read directly and uses no stationary law.

Two ways to attend less remove much of the sink mass. Replacing softmax with sigmoid attention in a CIFAR-100 ViT, with unnormalized scores and the output scaled by $1/N$, lets rows abstain (its sink readouts use the row-renormalized attention): the mean row mass falls to 0.34 at depth 12 and 0.31 at depth 24. Content-sink mass falls from 0.026 to 0.002 and from 0.033 to 0.001 (lower in 3/3 seeds at depth 12 and 3/3 at depth 24), and the share of layers with a sink from 0.53 to 0.22 and from 0.65 to 0.07 (lower in 2/3 seeds at depth 12, with 1 unchanged, and 3/3 at depth 24). Test accuracy falls from 55.9% to 53.3% and from 55.8% to 54.1% (lower in 3/3 seeds at depth 12 and 3/3 at depth 24). Massive activations do not fall consistently: at depth 12 they rise in 3/3 seeds (34.8 to 54.0), and at depth 24 the means barely move (60.9 to 61.3). This sigmoid is not the $-\log N$ rule of Section 17.2. A headwise output gate (Qiu et al., 2025) is a diagonal gate on the softmax (Section 8.2), so it too lets a head attend less. In a hybrid causal character-level model with two softmax layers it more than halves the mean content-sink mass, from 0.091 to 0.041 (lower in 3/3 seeds, by 0.060, 0.047, 0.042), while first-token mass (0.038 to 0.040), massive activations (9.64 to 9.29) and perplexity (5.80 to 5.79) barely move.

Extra tokens remove less. Adding four registers (Darcet et al., 2024) to a causal character-level WikiText language model lowers content-sink mass from 0.106 to 0.077 (change −0.029, 95% interval [−0.041, −0.017], five seeds) with no detectable change in its per-character perplexity (4.824 to 4.831;

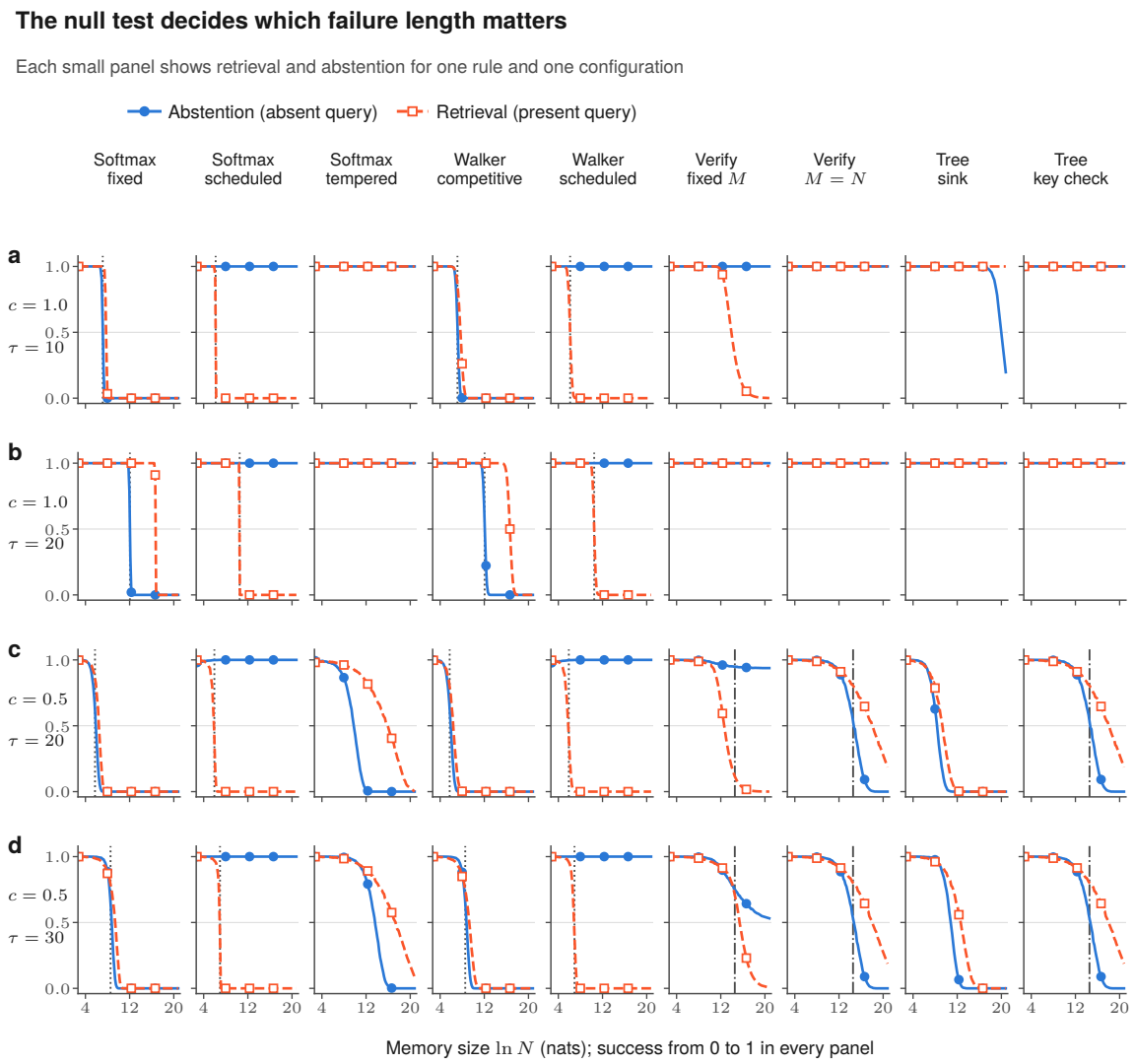


Figure 8.1: **A scheduled competitive null preserves abstention at a retrieval cost; absolute tests have a separate evidence limit.** Recovered product keys ($D = 128$), in four row configurations: (a,b) $c = 1$, $\tau = 10, 20$; (c,d) $c = 0.5$, $\tau = 20, 30$. All nine rule columns show absent-query abstention (blue filled circles, solid) and present-query retrieval (orange open squares, dashed). Nulls are calibrated at $N = 64$; walkers use $M = 64$ except the scaled verifier ($M = N$). Dotted lines show the fixed or scheduled competitive law; dash-dot lines show the saved mean-maximum threshold crossing, when finite. Tempering scales content and sink logits by $\ln N / \ln 64$; the tree sink competes over 4^2 shortlisted keys. Each point pools three seeds with 2,000 queries per type; conditional walker probabilities are exact. No seed intervals were retained. The saved execution failure remains; publication is a registered rebaseline, not historical reproduction (Bruce, 2026e, Appendix C.1).

Change	Model	n	Sink readout	Massive activation	Quality
Sigmoid, depth 12	CIFAR-100 ViT	3	$0.026 \to 0.002$ (3/3)	$34.8 \to 54.0$ (0/3)	$55.9\% \to 53.3\%$ (3/3)
Sigmoid, depth 24	CIFAR-100 ViT	3	$0.033 \to 0.001$ (3/3)	$60.9 \to 61.3$ (2/3)	$55.8\% \to 54.1\%$ (3/3)
Output gate	char-level LM	3	$0.091 \to 0.041$ (3/3)	$9.64 \to 9.29$ (1/3)	$5.80 \to 5.79$ (2/3)
Four registers	WikiText LM	5	$0.106 \to 0.077$ -0.029 [-0.041, -0.017]	$12.25 \to 11.31$ -0.94 [-2.07, 0.50]	$4.824 \to 4.831$ 0.008 [-0.031, 0.046]
Four registers	CIFAR-100 ViT	5	$0.0198 \to 0.0106$ -0.0092 [-0.0183, 0.0034]	$29.8 \to 120.3$ 90.5 [61.1, 132.7]	$56.77\% \to 56.32\%$ -0.45 [-0.71, -0.23]
Value-null register	byte-level LM	5	$0.0889 \to 0.0763$ -0.0126 [-0.0196, -0.0070]	$8.05 \to 8.05$ 0.00 [-0.41, 0.40]	$2.0231 \to 2.0201$ -0.0030 [-0.0064, -0.0002]
Anchor gate	76.7M decoder	2	first token: $0.0111 \to 0.0068$ (2/2)	—	$25.80 \to 26.85$ (0/2) at 2,048: $88.1 \to 59.8$ (2/2)
Observational	GPT-2 Medium	384	$\rho = -0.37$ [-0.46, -0.28]	—	—
Observational	Pythia-1.4B	384	$\rho = -0.45$ [-0.56, -0.34]	—	—
Observational	Llama-2-7B	1,024	$\rho = -0.50$ [-0.56, -0.45]	—	—

Table 8.1: Trained-model evidence on the sum-to-one constraint. Each entry shows the control mean $\to$ the mean after the change; n counts seeds, or heads for the observational rows. With five seeds the paired change and its 95% interval follow; with two or three seeds, the number of seeds in which the readout is lower after the change, and every seed's value is in Table 8.2. No p-values. The sink readout is content-sink mass, except first-token mass for the anchor gate and, for the observational rows, ρ below. Quality is test accuracy for the ViTs, bits per byte for the value-null register, and perplexity otherwise; the anchor gate's second entry is its perplexity at 2,048 tokens, four times its training length. The sigmoid is unnormalized with its output scaled by $1/N$. Content-sink mass in the causal models (the output gate, the registers' language model and the value-null register) rests on the value-norm test and a degenerate loading (see text). The observational rows are correlations: ρ is the Spearman correlation between a head's total sink mass and its share of the summed per-head output norms (before the output projection), and its intervals, from resampling heads of one model, are descriptive. The anchor gate's legacy LAMBADA scorer does not measure accuracy and is omitted. Production-ViT evidence, observational and with its own sink definitions, is in Table 8.3.

$[-0.031, 0.046]$). Its massive activations go from 12.25 to 11.31, but the interval $[-2.07, 0.50]$ spans zero, so we claim no reduction. In a CIFAR-100 ViT the same change raises massive activations in every seed, from 29.8 to 120.3 (change 90.5, $[61.1, 132.7]$), changes accuracy by -0.45 points ($[-0.71, -0.23]$), and leaves a sink-mass interval that spans zero ($[-0.0183, 0.0034]$). A register with a null value would be a query-dependent null logit (Proposition 8.2); in the test below, the register's embedding is certified to be annihilated by every value map. In a causal byte-level model it cuts content-sink mass by only 14% (0.089 to 0.076; $[-0.020, -0.007]$, five seeds), against the 40% its prespecified test asked for, and not detectably more than a plain register token (certified minus plain -0.003, $[-0.010, 0.006]$). The register receives 0.054 of the attention, its own row included; bits per byte improve slightly ($[-0.0064, -0.0002]$), and massive activations show no detectable change. Its dial is a learned scalar per layer, added to every query's attention logit for the register. Setting the dial moves the register's mass monotonically (Spearman 0.987 across six dial settings in

five seeds), but the median per-layer slope of the register mass's log-odds against the dial is 0.38 ($R^2 = 0.95$), below the prespecified band of 0.7 to 1.3. That mass includes the register's own row, a floor of 1/129, which flattens the slope at low settings.

An anchor trades quality for length. A per-head anchor gate mixes each head's output with the token's own value, $o = (1 - a)PV + aV$, so its rows still sum to one. In a 76.7M-parameter RoPE decoder trained at 512 tokens it cuts perplexity at 2,048 tokens from 88.1 to 59.8 (lower in 2/2 seeds), but raises it at the training length from 25.80 to 26.85 (higher in 2/2 seeds), and it lowers the first token's mass from 0.011 to 0.007.

Sink mass and content energy. Observationally, heads that put more mass on sinks carry less of their layer's output. Across the heads of GPT-2 Medium, Pythia-1.4B and Llama-2-7B, a head's total sink mass and its share of the summed per-head output norms (before the output projection) have Spearman correlations -0.37, -0.45 and -0.50 over 384, 384 and 1,024 heads. The sink mass is the causal detector's, so it rests on the value-norm test and a degenerate loading, and intervals from resampling heads are descriptive only, since heads of one model are not independent. A prespecified within-head test asked about a sharper link: whether a head's sink mass tracks its contraction defect beyond the mixing term. In this test, sink mass is the mean first-token weight over the head's content queries, with first-token values idealized as null because their value contribution is unavailable. The result was negative: 0 of 7 public causal language models show a positive association, with decision statistics from -0.175 to -0.083 (Bruce, 2026c, Appendix D).

Sink mass as an eviction rule. A prespecified test turned absorbed mass into a cache policy: on Pythia-1.4B, over coherent 2,048-token PG19 windows, it ranked cached tokens by how much of their content had drained into the sink through the layers and evicted the most absorbed first, always keeping the first token, the sink. Its retention certificate, a worst-case bound on how far the evictions can move the final-layer states, held with no violations and was non-vacuous on every window: it stayed inside the trivial bound. As a policy it lost. The policy comparison uses a linear model of the forward pass built from each layer's captured, head-averaged attention. At a 12.5% cache, the policy's mean relative deviation from the full cache's final-layer states exceeded that of H2O (Zhang et al., 2023), TOVA (Oren et al., 2024) and StreamingLLM (Xiao et al., 2024) by 0.018, 0.016 and 0.014. Matching H2O's deviation at a 14.0% cache took a 37.5% cache.

What the records show. Two ways to attend less, sigmoid attention and an output gate, remove much of the sink mass; sigmoid pays for it in accuracy. Extra tokens remove less. Registers lower sink mass in a language model but raise a ViT's massive activations, and a register whose embedding is certified value-null removes only 14%, not detectably more than a plain one. These are small models with few seeds and one change each, and none of them tests the two-sided law of Chapter 9. The trained harness was built to, but at 125M no retrieval head reached the registered identification threshold (Section 17.4). Its sink readouts agree with the small models: every rule that gives attention a null removed the first-token sink and cut massive activations, while a signed gate without a null did not (Figure 8.2).

Production vision transformers. Most of the measurements above come from small models trained for the purpose. An observational census of seven pretrained ViTs gives a check at production scale: DeiT-B, DeiT3-B, ViT-B pretrained on ImageNet-1k, ViT-B with its original ImageNet-21k pretraining, and ViT-S, ViT-B and ViT-L with AugReg ImageNet-21k pretraining, each measured on the same 1000 ImageNet validation images and re-analysed from the census records (Tables 8.3 and 8.4 and Figure 8.3; Bruce, 2026c). The census has two sink definitions, which we keep apart. A CLS-readout sink (sink_cls) is a patch token in the top quartile of the CLS token's attention and the bottom quartile of composed value norm. A π-outlier sink (sink_H) is a patch token whose stationary mass under the head-averaged patch chain is an outlier while its value norm and cluster loading are low. A bidirectional ViT has no causally absorbing token, yet 6 of the 7 checkpoints

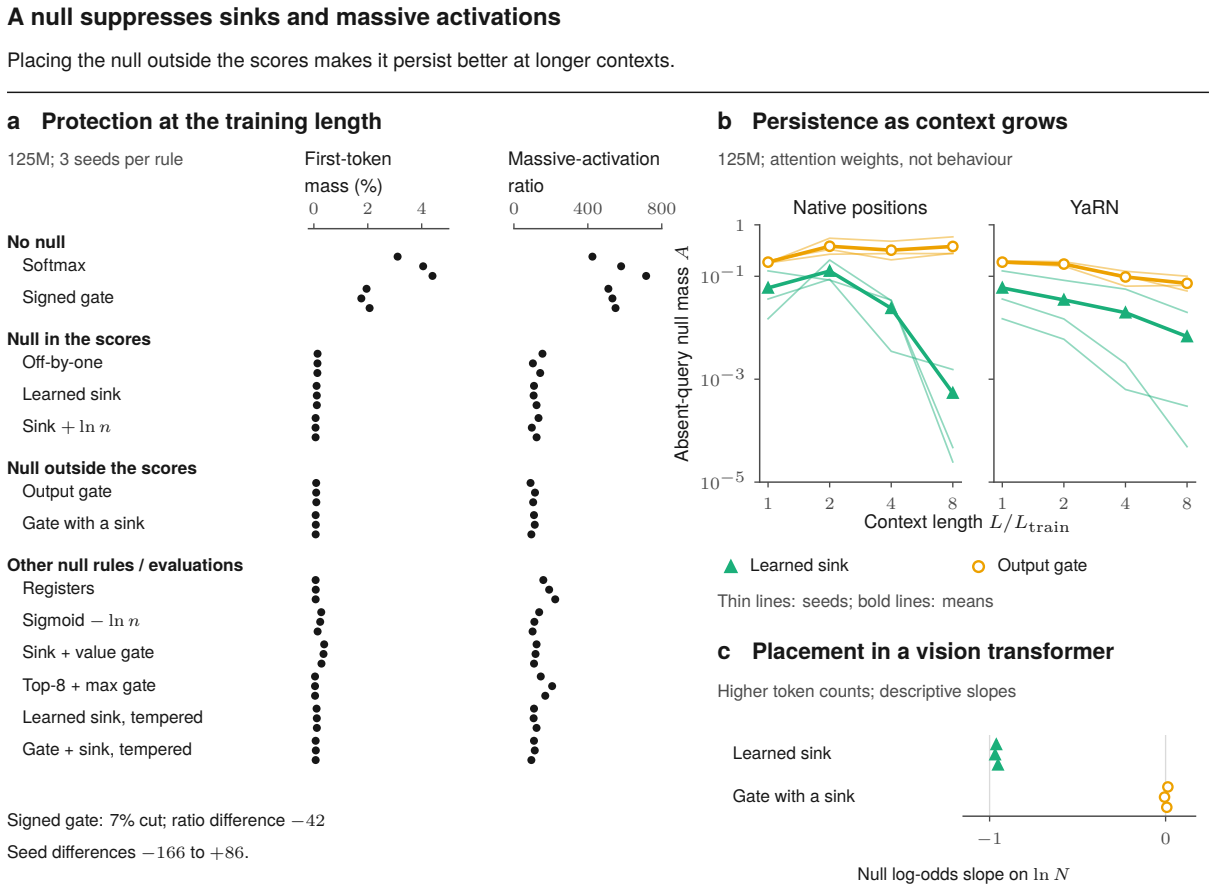


Figure 8.2: A null suppresses first-token sinks and massive activations in these language models, and its placement governs persistence at longer contexts. (a) All 13 evaluation variants of the 125M harness; each dot is one of 3 seeds, trained at 1,024 tokens. First-token mass is measured on FineWeb; the massive-activation ratio is the largest layerwise maximum-to-median absolute activation ratio. The registered comparison of the output gate and tempered gated sink held in language models but failed in vision transformers; the all-rule comparison is descriptive. The signed gate, which has no null, cuts the seed-mean ratio by 7%; its paired differences from softmax across the 3 seeds range from −166 to +86 (mean −42). Tempered evaluations reuse the learned-sink and gated-sink checkpoints and coincide with them at training length. (b) Descriptive absent-query null mass for the learned sink and output gate with native positions or YaRN; thin lines show seeds and bold lines their arithmetic mean. No retrieval head reached the identification threshold, so these are attention weights, not behavioural abstention. (c) Descriptive, unscored null log-odds slopes for one ViT size, 3 checkpoints per rule, on CLS rows of 1,000 ImageNet-100 validation images at 224–896 px, using upsampled crops. Each dot is a checkpoint summary: median across images for each head, then heads and layers; N is the patch count. The reference slopes −1 and 0 denote stationary dilution and unchanged odds, respectively.

grow a sink_H population, born between layers 6 and 11 with peak median counts of 3–10 tokens out of 196. Birth is the first layer at which at least half the images carry a flagged token. DeiT3-B grows none; instead its CLS token receives 11–36% of head-averaged attention across layers. Where a population exists, 87–99% of its tokens lie inside the CLS-readout corner. How long the sink_H set persists across depth differs between checkpoints (Figure 8.3a). The set locks at the first layer whose pooled Jaccard overlap with the next layer reaches 0.7, and dissolves at the first layer after its peak whose overlap with the layer before falls below one half. Its largest adjacent-layer Jaccard overlap is 0.074 in DeiT3-B, which has no population and few flagged tokens, and 0.97 in ViT-L; ViT-L locks, turns over across layers 17–19, locks again and dissolves at layer 22. Neither a set that re-forms at every layer nor a fixed register describes these models, and with one checkpoint per setting the difference between them cannot be attributed to the training recipe, the data or the scale. The census registered its predictions only for ViT-B with its original ImageNet-21k pretraining; the one about persistence, that the set locks and then dissolves, misses the lock rule at its point estimate (a largest Jaccard overlap of 0.688 against 0.7), although its interval spans the threshold. These in-sequence sinks are not value-null. Zeroing a sink token's value at one layer, with attention unchanged, moves the mean-centred output logits 1.5–1.9 times as much as zeroing a typical patch for sink_cls, and 2.4–10.6 times as much for sink_H, in all 6 checkpoints with ground truth; removing the key adds a renormalization effect on top. sink_H delivers latest in depth (transport centre of mass at 0.73–0.86 of depth), partly because the set exists only from its birth layer on. Both populations carry content of their own as well as holding attention that would otherwise go to other keys, so the value-null factorization of Section 8.2 does not describe them. The census detector differs from the one behind Table 8.1, so its counts are not compared with that table's.

8.6 Own-share floors

Abstention returns in test-time context mixing, where, for an affine readout and unit row sums, every margin splits exactly into a sample's own contribution and its companions'. If row mass is lost, an affine-bias correction is also needed (Section 12.5). A head that can abstain makes that mixing sub-stochastic, and the questions it raises there are the loop gain of a sub-stochastic softmax, the floor on a sample's own share at collapse, and the failure of a fixed null. At collapse, a fixed null competes with near-equal keys, so only a null coupled to the temperature, or gated by the query, should repair abstention across temperatures with one shared setting. Chapter 0 records this as a prediction not yet run (Section 0.4.6); its full treatment is left for later work.

8.7 What holds in trained networks

Exact for every network of the stated class. Sum-to-one and the gate-times-softmax form (Sections 8.1 and 8.2): unit row sums, the null direction of the softmax Jacobian and the gate-times-softmax factorization hold for every head, at every layer and input. Rollout as survival, and the order of the sink treatments (Section 8.3): exact for ordered products of causal stochastic kernels (the first token absorbing) with per-token residual shares, whatever the layers: layer-varying by construction. It describes the rollout linearization, not the full block.

Measured in trained networks. Sinks in production vision transformers (Section 8.5): observational measurement in pretrained models.

Change	Readout	Arm	Seed 0	Seed 1	Seed 2
Sigmoid, depth 12	content-sink mass	softmax	0.01893	0.0485	0.01132
		sigmoid	0.00144	0.004878	0
	layers with a sink	softmax	0.6667	0.6667	0.25
		sigmoid	0.3333	0.08333	0.25
	row mass	softmax	1	1	1
		sigmoid	0.3494	0.3258	0.3493
	massive activation	softmax	39.61	36.64	28.05
		sigmoid	43.29	68.85	49.92
	test accuracy (%)	softmax	55.54	55.92	56.18
		sigmoid	53.17	52.96	53.9
Sigmoid, depth 24	content-sink mass	softmax	0.03982	0.02232	0.03661
		sigmoid	0.002755	0.000698	0.0008294
	layers with a sink	softmax	0.625	0.5833	0.75
		sigmoid	0.08333	0.08333	0.04167
	row mass	softmax	1	1	1
		sigmoid	0.3211	0.3185	0.302
	massive activation	softmax	80.04	51.68	50.9
		sigmoid	65.02	70.89	47.9
	test accuracy (%)	softmax	55.31	56.18	56.01
		sigmoid	54.08	54.23	54.02
Output gate	content-sink mass	ungated	0.1087	0.06659	0.09815
		output gate	0.04826	0.01948	0.05598
	first-token mass	ungated	0.03589	0.03749	0.04167
		output gate	0.03793	0.03322	0.04862
	massive activation	ungated	8.026	11.68	9.215
		output gate	8.662	9.861	9.347
	perplexity	ungated	5.826	5.765	5.798
		output gate	5.807	5.803	5.761
Anchor gate	first-token mass	baseline	0.01116	0.01112	
		anchor gate	0.006846	0.006812	
	perplexity, 512 tokens	baseline	25.72	25.88	
		anchor gate	26.96	26.74	
	perplexity, 2,048 tokens	baseline	86.71	89.52	
		anchor gate	61.12	58.53	

Table 8.2: Every seed's value behind the comparisons of Table 8.1 with fewer than five seeds: sigmoid attention at depths 12 and 24 (softmax and sigmoid arms), the output gate (ungated and gated) and the anchor gate (baseline and gated). Content-sink mass for the output gate rests on the value-norm test and a degenerate loading; the anchor gate's first-token mass uses no stationary law.

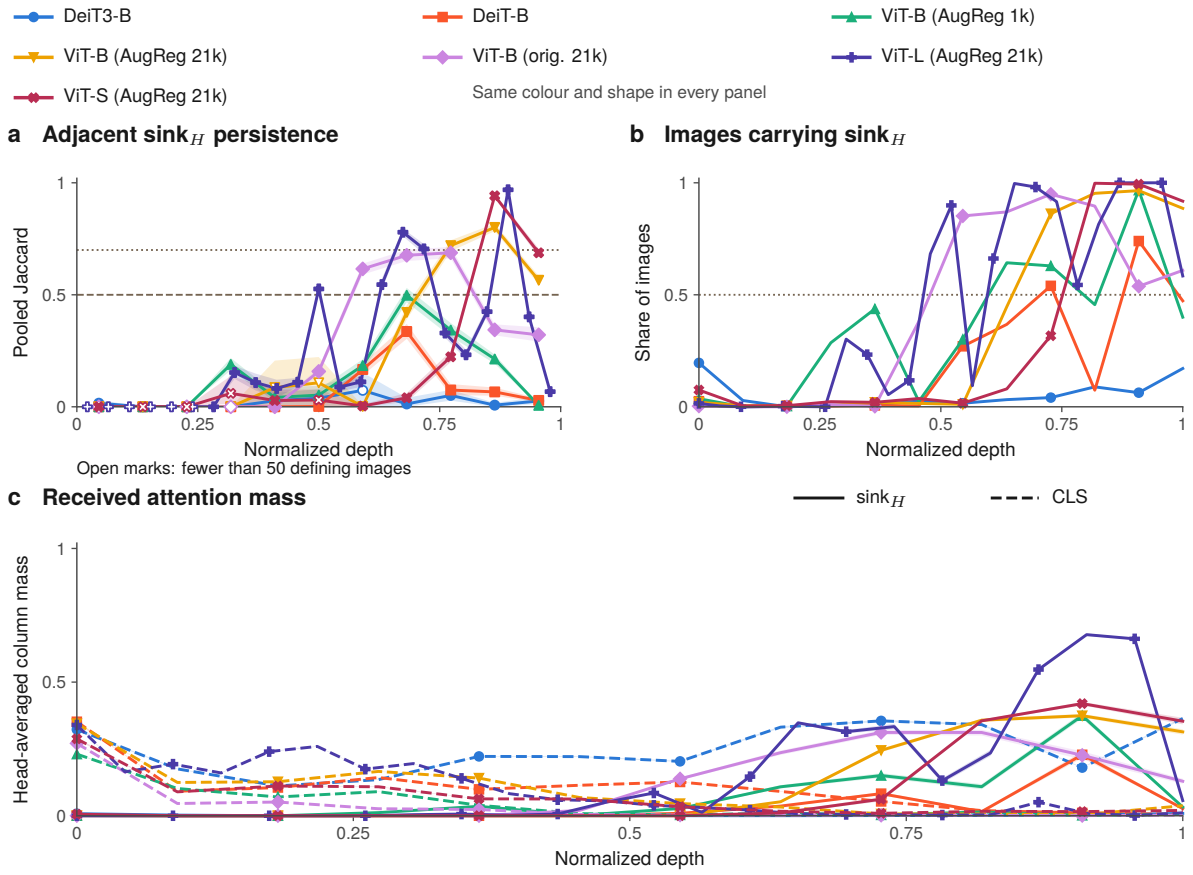


Figure 8.3: Sink persistence differs across the seven pretrained ViT checkpoints. (a) Pooled adjacent-layer Jaccard of the sink_H set, with hollow model markers where fewer than 50 images define the pair; dotted and dashed rules mark 0.7 and 0.5. (b) Share of images carrying sink_H; the dotted rule marks 0.5. (c) Head-averaged received attention mass: solid sink_H, dashed CLS. Checkpoint colours and marker shapes are fixed across panels. Bands show the recorded 95% cluster-bootstrap intervals over images: 400 resamples for Jaccard and 2,000 for census mass. 1,000 images per checkpoint. Patch tokens only for sink_H (the π-outlier set); sink_cls is the CLS-readout corner, whereas CLS mass is the CLS column itself. The lifecycle prediction is registered for ViT-B original-21k only; other checkpoints are robustness evidence. One checkpoint per setting cannot separate recipe, data and scale. The detector differs from the one behind Table 8.1, and the counts are not compared. Records: the census records (Bruce, 2026c).

Lifecycle of the sink_H set

Checkpoint	Blocks	Birth	Peak (median)	Largest image share	Lock	Dissolution
DeiT3-B	12	—	—	0.196 (L0)	—	—
DeiT-B	12	8	L10: 3	0.740 (L10)	—	11
ViT-B (AugReg 1k)	12	7	L7: 3	0.967 (L10)	—	8
ViT-B (AugReg 21k)	12	8	L8: 4	0.965 (L10)	8	—
ViT-B (orig. 21k)	12	6	L8: 9	0.949 (L8)	—	10
ViT-L (AugReg 21k)	24	11	L20: 10	1.000 (L20)	15	22
ViT-S (AugReg 21k)	12	9	L9: 5	0.998 (L9)	9	—

Persistence and overlap with the CLS-readout corner

Checkpoint	J max (pair)	95% CI	Images	H inside C	C that is H
DeiT3-B	0.074 (L6→7)	[0.000, 0.153]	40	0.519	0.009
DeiT-B	0.336 (L7→8)	[0.303, 0.370]	643	0.989	0.063
ViT-B (AugReg 1k)	0.498 (L7→8)	[0.467, 0.531]	807	0.872	0.073
ViT-B (AugReg 21k)	0.801 (L9→10)	[0.784, 0.818]	987	0.966	0.113
ViT-B (orig. 21k)	0.688 (L8→9)	[0.663, 0.714]	987	0.971	0.253
ViT-L (AugReg 21k)	0.969 (L20→21)	[0.963, 0.974]	1000	0.891	0.212
ViT-S (AugReg 21k)	0.942 (L9→10)	[0.931, 0.952]	999	0.956	0.086

Table 8.3: The π-outlier sink set (sink_H) in seven pretrained ViTs: its lifecycle across depth, how it persists, and how it overlaps the CLS-readout corner (sink_cls). Birth is the first layer at which at least half the images carry sink_H; the peak is the layer with the largest median count, given with that count; the largest image share is the largest share of images carrying sink_H, with its layer. Lock is the first layer whose set overlaps the next layer's with a Jaccard index of at least 0.7, pooled over images, and dissolution is the first layer after the peak whose overlap with the layer before falls below 0.5; zero-median trajectories have no peak or inferred dissolution. J max is the largest adjacent-layer overlap, with its layer pair, its 95% interval and the number of images that define it. The last two columns give the share of sink_H tokens inside sink_cls and the share of sink_cls tokens that are sink_H. 1,000 images per checkpoint; head-averaged; patch tokens only. Intervals: 400-resample cluster bootstrap over images (Bruce, 2026c).

One-layer counterfactual ratios against a typical patch: ratio (Cliff's δ)

Checkpoint	sink_H transport	sink_H modulation	sink_cls transport	sink_cls modulation
DeiT3-B	3.84 (0.60)	6.46 (0.66)	1.54 (0.41)	1.57 (0.46)
DeiT-B	5.17 (0.84)	10.05 (0.92)	1.60 (0.50)	1.60 (0.52)
ViT-B (AugReg 1k)	5.16 (0.94)	13.34 (0.97)	1.88 (0.56)	2.28 (0.69)
ViT-B (AugReg 21k)	10.65 (0.90)	29.46 (0.99)	1.74 (0.58)	1.86 (0.62)
ViT-B (orig. 21k)	2.35 (0.56)	6.29 (0.85)	1.65 (0.47)	2.14 (0.60)
ViT-L (AugReg 21k)	4.63 (0.91)	8.42 (0.96)	1.62 (0.53)	1.86 (0.61)
ViT-S (AugReg 21k)	—	—	—	—

Delivery depth and attention mass

Checkpoint	sink_H depth COM	sink_cls depth COM	Luminous depth COM	Peak sink_H attention mass	95% CI	CLS mass, second half
DeiT3-B	0.727	0.598	0.537	0.009 (L11)	[0.007, 0.011]	0.296
DeiT-B	0.810	0.697	0.558	0.228 (L10)	[0.216, 0.240]	0.051
ViT-B (AugReg 1k)	0.748	0.541	0.576	0.374 (L10)	[0.364, 0.386]	0.006
ViT-B (AugReg 21k)	0.862	0.627	0.600	0.374 (L10)	[0.367, 0.381]	0.022
ViT-B (orig. 21k)	0.751	0.523	0.463	0.312 (L8)	[0.303, 0.321]	0.002
ViT-L (AugReg 21k)	0.797	0.615	0.544	0.677 (L21)	[0.673, 0.682]	0.015
ViT-S (AugReg 21k)	—	—	—	0.420 (L10)	[0.413, 0.426]	0.019

Table 8.4: The value channels and delivery depth of the two sink populations. Editing one layer, transport zeroes a token's value with attention unchanged; modulation is the renormalization effect of deleting the token from the key set, the attention mass it held spreading over the other keys. Each entry is the median over layers of a per-layer median over images: the population's effect on the mean-centred output logits divided by the all-patch median, with Cliff's δ against the other patch tokens in parentheses. DeiT3-B's sink_H ratios use only two layers, and ViT-S has no counterfactual ground truth. Depth COM is the centre of mass of a population's transport over the layers, as a fraction of depth; sink_H's is late partly because the set exists only from its birth layer on. Luminous tokens lie in the top quartile of both the CLS token's attention and composed content. The peak sink_H attention mass is the largest, over layers, of the mean head-averaged attention the set receives, with its layer and 95% interval (2,000-resample bootstrap over images); it is a size statistic partly induced by selection. The last column is the CLS token's mass averaged over the second half of depth. 1,000 images per checkpoint; head-averaged; patch tokens only (Bruce, 2026c).

Chapter 9

A margin that grows with length: the routing budget

The fifth condition of Chapter 0 is a routing budget (Section 0.3.2). The more tokens compete, the larger the score advantage the right one needs: one key among N needs a logit margin that grows like $\log N$, so with a bounded logit scale retrieval fades as the context grows. Once a head has a null, the way to deliver nothing of Chapter 8, the margin has two sides, because abstention needs the null to outweigh the background, and retrieval needs the match to outweigh the null and the background together. Section 9.1 makes this budget exact, and Section 9.2 gives its law as the context grows, with the rule that keeps both sides. Section 9.3 splits the budget into two, a budget of evidence, which belongs to the keys, and a budget of gain, which belongs to the read; Section 9.4 shows that, with an exchangeable background, decisions on sums need a $\log N$ schedule at a fixed logit scale while decisions on maxima need none against backgrounds bounded below their threshold; and Section 9.5 prices naming one of N in every machine's currency. Section 9.6 measures the retrieval half on length and resolution, and Section 9.7 says what of this holds in trained networks. Chapter 17 applies the theorems to abstention and retrieval at long context, with the trained evidence for both sides.

9.1 The exact simultaneous budget

Compare two matched queries. The *absent* query sees N background keys with scores $s_1, \dots, s_N$, total background weight $B_N = \sum_j e^{s_j}$, and a null of weight $z_N > 0$, one of the nulls of Chapter 8 (a single sink logit b has $z_N = e^b$; several value-null registers count through their summed weight). The *present* query sees the same background and null plus one matching key with score s_*. Abstention and retrieval are

$$a_N = \frac{z_N}{z_N + B_N}, \qquad w_N = \frac{e^{s_*}}{e^{s_*} + z_N + B_N}.$$

The matched convention is what makes the claim joint: the null cannot know whether the match is present. A gate that detects presence from outside the normalization is outside this class. Write $A_* = a_*/(1-a_*)$ and $W_* = w_*/(1-w_*)$ for the odds of two targets. That a winning score's margin must grow like $\log N$ for retrieval alone is known (Nakanishi, 2025; Veličković et al., 2025); the theorem adds the abstention side and the joint budget.

Theorem 9.1 (Exact simultaneous budget)**.** *For $0 < a_*, w_* < 1$, both $a_N \geq a_*$ and $w_N \geq w_*$ hold if and only if*

$$A_* B_N \;\leq\; z_N \;\leq\; e^{s_*}/W_* - B_N,$$

and such a null exists if and only if $e^{s_*} \geq w_* B_N/((1-w_*)(1-a_*))$. *If every content score lies in* $[-\tau,\tau]$, *no choice of null offset, fixed, learned, or a learned offset plus* $\log N$, *satisfies both beyond*

$$N_* = e^{2\tau}(1-a_*)\frac{1-w_*}{w_*},$$

and the bound is sharp: distractors at $-\tau$, *the match at* τ *and* $z_N = A_* N e^{-\tau}$ *attain it.*

Proof. Clearing the positive denominators turns each target into one side of the interval. The interval is nonempty exactly when $(1+A_*)B_N \leq e^{s_*}/W_*$, which is the stated condition. With bounded scores, $B_N \geq Ne^{-\tau}$ and $e^{s_*} \leq e^{\tau}$; the construction attains equality. □

The budget is one inequality with the null on both sides of it. Abstention needs the null to outweigh the background; retrieval needs the match to outweigh the null and the background together. Growing the background by a factor therefore squeezes the null's window from both ends, and with a bounded logit scale the window closes at N_*.

The budget binds because the matched queries share one null. A null that may differ between them by a bounded amount pays a known price.

Proposition 9.2 (Presence-sensitive null)**.** *Let the absent and present queries have background weights* $B_0, B_1 > 0$, *let the match have weight* $M > 0$, *and let their nulls have weights* $z_0, z_1 > 0$ *with* $|\log z_0 - \log z_1| \leq \epsilon$. *Both targets can be met if and only if*

$$M \;\geq\; W_*(B_1 + A_* e^{-\epsilon} B_0).$$

If every content score lies in $[-\tau,\tau]$, *no such nulls meet both targets beyond*

$$N_\epsilon = e^{2\tau}\,\frac{1-w_*}{w_*\,(1+A_* e^{-\epsilon})}$$

background keys, which rises from N_* *at* $\epsilon = 0$ *to* $N_*/(1-a_*)$ *as* $\epsilon \to \infty$, *and the bound is sharp. If* B *and* M *are fixed with* $A_* B > M/W_* - B$, *a random null that meets the absent target with probability at least* $1-\alpha$ *and the present one with probability at least* $1-\beta$ *has absent and present laws that differ by at least* $1-\alpha-\beta$ *in total variation.*

The proof is in Section A.10. Sensitivity to presence can remove at most the abstention penalty, the factor $1-a_*$ in N_*. It cannot remove the dilution of the match's weight by the background. It also needs evidence of presence: whatever the null is computed from must tell the two queries apart about as reliably as the targets demand.

The budget says what each rule does. A fixed sink, $z_N = e^b$, gives

$$a_N \leq \sigma(b+\tau-\log N), \qquad w_N \leq \frac{e^{2\tau}}{e^{2\tau}+N},$$

with σ the logistic function, so both vanish; whenever the null scores below the match, as it must for the match to win, the abstention bound reaches any given level first, at $e^{b-\tau}$ times the length at which the retrieval bound does. A null calibrated to the length, $z_N = Ne^b$, holds abstention constant against a background of equal scores but gives $w_N = O(1/N)$. Sigmoid weights behave the same way.

Proposition 9.3 (Sigmoid weights)**.** *Let* $p_j = \sigma(s_j + c_N)$ *with one offset* c_N *shared by the matched queries, and define abstention, when* $\sum_j p_j \leq 1$, *as* $1-\sum_j p_j$. *If* $|s_j|, |s_*| \leq \tau$, *the absent row mass is at most* $M = 1-a_*$, *and the match weight is at least* w_*, *then necessarily* $N \leq M\big(1+e^{2\tau}(1-w_*)/w_*\big)$.

The proof is in Section A.10. A $-\log N$ bias is the choice $c_N = c - \log N$ that holds the row mass fixed; the proposition shows that it pays with single-key retrieval, whose weight then falls like $1/N$. None of these statements rules out perfect classification from a shrinking but nonzero signal in exact arithmetic: they bound attention weight.

Two further statements matter when these laws are compared with measurements. If N counts all keys of the present query rather than the distractors, replace N by $N-1$ in the retrieval statements; the trained pilot of Section 17.2, whose onsets Figure 9.2 plots, counts this way, since its present rows have $N-1$ distractors. And a realized abstention at a finite length fluctuates about its mean-field value: under the tempering of Section 9.2, which multiplies every score, the null logit b included, by $\lambda_N = \log N / \log N_0$ for a head trained at N_0 keys, the almost-sure converse of that section (Theorem 9.8) has a finite-sample version.

Proposition 9.4 (Finite-sample comparison)**.** *In the bounded i.i.d. setting let* $U_{j,N} = e^{-\lambda_N(s_0 - S_j)} \in [0,1]$, $\mu_N = \mathbb{E}U_{j,N}$ *and* $v = N_0 e^{s_0 - b}$. *For* $0 < \delta < 1$ *and* $t_N = \sqrt{\log(2/\delta)/(2N)}$, *with probability at least* $1-\delta$,

$$\frac{1}{1 + v^{\lambda_N}(\mu_N + t_N)} \le a_N \le \frac{1}{1 + v^{\lambda_N}\max(0, \mu_N - t_N)},$$

and $\mathbb{E}a_N \ge \left[1 + v^{\lambda_N}\mu_N\right]^{-1}$.

The proof is in Section A.10. Here the background scores S_j are i.i.d. and bounded above, with essential supremum s_0, as in Theorem 9.8, and v is the abstention margin v_a of Theorem 9.6.

Without a null the budget is retrieval alone. A single key competing with N others at a common logit needs the margin $\log(N(1-\eta)/\eta)$ to keep a share $1-\eta$, and content or noise in the keys adds to it: that is the restricted capacity theorem of Section 6.1 (Theorem 6.8), which prices address margin, content scale, noise, width and readout gain separately. At head width no larger than the content dimension, when content must be recovered through the keys, the budget is also dimensional (Section 6.1). Section 9.6 measures this retrieval half on length and resolution.

9.2 The mean-field law and tempering

When the background exponentials e^{s_j} are i.i.d. with mean μ, abstention follows the mean-field prediction $a_{\mathrm{mf}} = z_N/(z_N + N\mu)$, and a fixed sink crosses one half at $N_{1/2} = e^b/\mu$.

Proposition 9.5 (Concentration and expectation error)**.** *If* $e^{s_j} \in [l,u]$ *with* $l > 0$, *then with probability at least* $1-\delta$,

$$\frac{z_N}{z_N + N(\mu + t)} \le a_N \le \frac{z_N}{z_N + N\max(l, \mu - t)}, \qquad t = (u-l)\sqrt{\log(2/\delta)/(2N)},$$

and $0 \le \mathbb{E}a_N - a_{\mathrm{mf}} \le z_N N v/(z_N + Nl)^3$, *where* v *is the variance of* e^{s_j}.

The proof is in Section A.10. For Gaussian logits $s \sim \mathcal{N}(m, v_s)$, $\mu = e^{m + v_s/2}$, and after tempering by λ, below, the background moment is $e^{\lambda m + \lambda^2 v_s/2}$; the untempered μ is wrong after scaling. The moment predicts the realized background sum only while $\lambda\sqrt{v_s} < \sqrt{2\log N}$; past that scale the sum is carried by its few largest terms, and the moment overstates it by a factor of order $e^{(\lambda\sqrt{v_s} - \sqrt{2\log N})^2/2}$ or more. A set of k near-matches contributes its exact sum separately. Heavy-tailed logits may have $\mathbb{E}e^s = \infty$ and then no finite mean-field law, while Theorem 9.1 still holds with the measured log-sum-exp. In the pilot of Section 17.2, the mean-field law, evaluated exactly with each head's learned logits, tracks the trained fixed-sink heads to within 0.002 per seed in all four configurations.

Suppose the head was trained at N_0 keys and is used at $N = N_0^\lambda$, that is $\lambda = \log N / \log N_0$, and multiply every score, the null included, by λ. This tempers the whole budget. The ratios in the theorem compare the largest background sum allowed at the training length with the null and with the match, and the null with the match.

Theorem 9.6 (Sufficient separated margins)**.** *Let the background scores be at most s_0, with $s_* > b > s_0$ fixed before scaling, and put $v_a = N_0 e^{s_0 - b}$, $v_r = N_0 e^{s_0 - s_*}$ and $u_r = e^{b - s_*}$. After tempering,*

$$a_N \geq \left(1 + v_a^\lambda\right)^{-1}, \qquad w_N \geq \left(1 + u_r^\lambda + v_r^\lambda\right)^{-1}.$$

If $v_a, v_r, u_r \leq 1$, both lower bounds are non-decreasing in N; if all three are below one, both tend to one.

The proof is in Section A.10. The margins are real conditions. If the null only equals the level of a uniform background, $b = s_0$, tempering changes no relative score of the absent query and $a_N = 1/(N+1)$ at every λ; more generally $b - s_0 < \log N_0$ makes the abstention bound decay even when the null outranks every single distractor. So a growing logit budget is necessary in this class (Theorem 9.1), and a particular schedule is not sufficient without the margins.

With one common multiplier, abstention depends exactly on a power mean of the exponentiated background scores. This average lies between their ordinary mean and their maximum.

Theorem 9.7 (Power-mean law)**.** *For an integer $N \geq N_0$, $\lambda = \log N / \log N_0$, and arbitrary finite scores, let $M_{q,N} = \left(N^{-1} \sum_j e^{q s_j}\right)^{1/q}$. Then exactly*

$$a_N = \left[1 + \left(N_0 e^{-b} M_{\lambda,N}\right)^\lambda\right]^{-1}, \qquad \frac{1}{1 + \left(N_0 e^{s_{\max} - b}\right)^\lambda} \leq a_N \leq \frac{1}{1 + \left(N_0 e^{-b} M_{1,N}\right)^\lambda}.$$

The proof is in Section A.10. The upper bound, which uses the ordinary mean $M_{1,N}$, can never certify a target; the lower bound is the separated-margin bound of Theorem 9.6. The law needs the same multiplier on every score. YaRN multiplies every content score by a temperature (Peng et al., 2024); a sink logit left out of that factor breaks the premise, and in an explicit example tempering then loses abstention (0.62 to 0.37 from one to eight times the training length) where the common multiplier keeps it (0.62 to 0.66). The trained harness gave its tempered sinks YaRN's factor as well, so there the premise held (Section 17.4). For a random background the law has a sharp phase boundary.

Theorem 9.8 (Almost-sure converse)**.** *Let the background scores be i.i.d. with essential supremum s_0 and $p_0 = \mathbb{P}(S = s_0)$, and let $h = \log N_0 + s_0$. Under tempering, almost surely*

$$a_N \longrightarrow \begin{cases} 0, & b < h, \\ (1 + p_0)^{-1}, & b = h, \\ 1, & b > h. \end{cases}$$

A present match with score s_* keeps weight tending to zero if $s_* < \max(b, h)$, and to $\left[1 + \mathbf{1}_{\{s_* = b\}} + p_0 \mathbf{1}_{\{s_* = h\}}\right]^{-1}$ otherwise. If the score law has unbounded upper support, $a_N \to 0$ for every finite b.

The proof is in Section A.10. So the abstention margin $v_a \leq 1$ is necessary as well as sufficient, up to the boundary. The boundary depends on the atom p_0 at the top of the score law, meaning the probability of a score exactly at that top. Here s_0 must be the essential supremum, the smallest upper bound that holds with probability one, not a loose bound. The same conclusions hold for stationary ergodic backgrounds; a nonergodic one need not use its unconditional supremum. At the boundary the rate of convergence is set by the upper tail, the probability of scores close to the top.

Theorem 9.9 (Boundary rate)**.** *If in addition* $\mathbb{P}(s_0 - S \le t) \sim Ct^a$ *as* $t \downarrow 0$*, with* $C > 0$ *and* $a > 0$*, then at* $v_a = 1$*, almost surely* $1 - a_N \sim C\,\Gamma(a+1)\lambda^{-a}$*, and a null lowered by* $(a \log \lambda - \log(C\Gamma(a+1)) + t)/\lambda$ *below the boundary holds abstention at* $(1 + e^t)^{-1}$*.*

The proof is in Section A.10. These results concern attention weights under a fixed score law. They do not identify the law a network learns as the context changes, and they do not turn an accuracy onset into a weight threshold; Proposition 9.4 is a finite-sample version of Theorem 9.8.

The theorems bound weight. The retrieval records of Section 9.6 measure accuracy under a learned decoder on noiseless values instead. Their native onsets, the first lengths at which that accuracy falls below one half, are exactly 8,192, 32,768 and 131,072 keys at training lengths 32, 64 and 128, and in the trained logit gap g_0 they follow $\log_2(\text{onset}) = 3.085\, g_0 + 0.003$, that is an onset growing as $e^{2.14\, g_0}$, faster than the weight bound's e^{g_0}. A learned decoder scored by nearest cosine can read a match whose weight has already fallen, so the accuracy onset is a separate estimand. A trained network can also escape a fixed sink's decay by lowering its background logits as the context grows. The sink's log-odds against the background, $y_i = b - \mathrm{LSE}_j(s_{ij})$, fall with slope $-\gamma$ in $\log i$, where γ is how fast the background grows; a stationary background with a finite exponential moment has $\gamma = 1$. In small trained models without positional encoding and with a learned null, the median within-model slope, pooled over the heads of 5 seeds at each training length, is -1.04 to -1.01 in the models trained at more than 64 tokens and -1.31 at 64 tokens (Section 17.4). With rotary positions the kernel holds the background down, and training widens it only when the null itself grows with length. The 125M results in this paragraph average each model's median head slope over 3 seeds per rule. On FineWeb-Edu text inside the training length, the background grows like $n^{0.12}$ for a fixed null, whose log-odds then fall at only -0.12, and like $n^{1.07}$ for a null that adds $\ln n$, so abstention stays nearly level under either rule (Section 17.4). Small byte models show the same contrast with the positions manipulated: γ is 0.12 to 0.19 with rotary positions and no null, an output gate or a fixed null, 0.97 with a null that adds $\ln n$, and 1.08 to 1.13 without positions whatever the null (Section 17.4; Figure 9.1). The compensation works through position and holds while positions stay in the trained range: beyond it the fixed sinks' slopes are -0.29 and -0.28 with positions interpolated (YaRN), and about -6.2 natively. The production model gpt-oss (OpenAI, 2025) has learned sink logits in dense layers and in 128-token banded layers, whose background stops growing at 128 keys, a built-in control. A prespecified test of its slopes was undecided by its rule: its books control failed, and the measured slopes match neither prespecified branch, though descriptively they lean toward compensation through position (Section 17.3).

The form of the compensation. Each head's content logits split exactly into a level r_i for each query, a popularity c_j for each key and a cycle part C_{ij}, the remainder orthogonal to every such sum: $s_{ij} = r_i + c_j + C_{ij}$, fitted by least squares over the sampled query rows and every key each one sees. The level and popularity together form the gauge part. Then $\mathrm{LSE}_j(s_{ij}) = r_i + \mathrm{LSE}_j(c_j) + X_i$ exactly, where X_i is the log of $e^{C_{ij}}$ averaged under the popularity weights, so the background's growth is a level term plus a popularity term plus a cycle term. A softmax over the keys cannot see the level, which cancels from its weights; the null's log-odds y_i see all three terms. In the 125M language models with a fixed null, inside the training length, the level's slope is -1.55 per e-fold of visible keys, the popularity term's 1.86 and the cycle term's -0.003 (95% interval $[-0.012, 0.011]$; medians over 432 heads, on eight PG-19 books). The pooled estimates here use 3 seeds at 125M and four scored seeds in the byte models; their intervals are 95% cluster bootstrap intervals, resampling seeds and then heads within each seed. So the compensation lives in the gauge part, which includes the linear part of a recency kernel: a bias linear in $i - j$ is exactly a level falling with i plus a popularity rising with j, with no cycle part. The softmax never sees the falling level, but against the null it offsets most of the popularity term's growth. The small byte models with rotary

positions and a fixed null show the same pair (−1.49 and 2.49) beside a real cycle term (−0.70, [−0.85, −0.57]); without positions the level is flat (−0.02) and the popularity term grows at 1.11 to 1.12, near the rate of 1 for keys of equal popularity. A null that adds $\ln n$ flattens the level's fall, by 1.12 per e-fold at 125M ([0.95, 1.27]) and 1.22 in the byte models: each model sets its gauge to the null it has. Beyond the training length the level reverses natively. On one split fitted over every length, its slope beyond the training length exceeds its slope inside it by 7.03 per e-fold at 125M ([6.53, 7.48]) and 1.89 in the byte models. Interpolated positions hold the background instead. At 125M with a fixed null, beyond the training length, the background grows like $n^{0.28}$ to $n^{0.35}$ across seeds under interpolation, against $n^{6.5}$ to $n^{7.3}$ natively, and interpolation holds it by steepening the gauge: the level falls faster and the popularity term grows faster than within the training length. A prespecified prediction that interpolation would keep the level's slope within 0.6 of its in-range value therefore failed at 125M (1.55, [1.41, 1.66]), while it held in the byte models (a shift of 0.31).

Figure 9.2 sets these laws against the trained pilot of Section 17.2: for every rule, logit budget and training length, the first length at which abstention or match weight falls below one half, against the onset its mean-field law predicts.

Figure 9.3 draws the three regimes on an exact construction: with fixed scores both weights eventually fall, adding $\log(N/N_0)$ to the null alone holds abstention while the match weight vanishes, and multiplying every score, the null included, by $\lambda_N = \log N/\log N_0$ keeps both where the three sufficient-margin factors of Theorem 9.6 are below one.

9.3 Two budgets: evidence and gain

A router must name one of N items from the scores it computes against them, and two different things can stop it. The keys may not carry enough evidence: some background item may score as high as the target. Or the read may not express what the keys carry: a soft router must put its mass on the target against N competitors. One limit belongs to the keys and the other to the read; in the language of decoding they are the code limit and the decoder limit (Shannon, 1959; Forney, 1968). Every machine's routing budget has both parts, and only the decoder limit depends on the machine. Write $s_1, \dots, s_N$ for the scores and t for the target.

Proposition 9.10 (Evidence cannot be bought)**.**

(a) If a soft router with weights proportional to $e^{\tau s_j}$, $\tau > 0$, puts at least half its mass on the target, even with an extra nonnegative null weight in its normalizer, then the target is the argmax: $s_t \geq \max_{j \neq t} s_j$.

(b) Let the target be uniform over the N items, and let the joint density of the scores, given that the target is t, factor as $p_t(s) = g(s)\,\ell(s_t)$ with ℓ nondecreasing. Independent background scores with density f_0 and an independent target score with density f_1, with f_1/f_0 nondecreasing, are an example. Then the argmax is the MAP rule: no rule that sees only the scores, randomized or not, finds the target more often.

(c) If the $N-1$ background scores are independent with continuous distribution function F and the target scores c, the target is the argmax with probability $F(c)^{N-1}$. If $0 < F(c) < 1$, its real-valued half-success crossing is $N_{\mathrm{ev}} = 1 + \ln 2/(-\ln F(c))$.

The proof is in Section A.10. Part (a) needs nothing. Part (b) needs its joint condition: exchangeable backgrounds with a monotone marginal likelihood ratio are not enough, and Section A.10 gives a three-item counterexample. Part (c) is the argmax's exact curve for a fixed target score, not a ceiling by itself, since a rule that knew c exactly could pick the item that scores c. For random

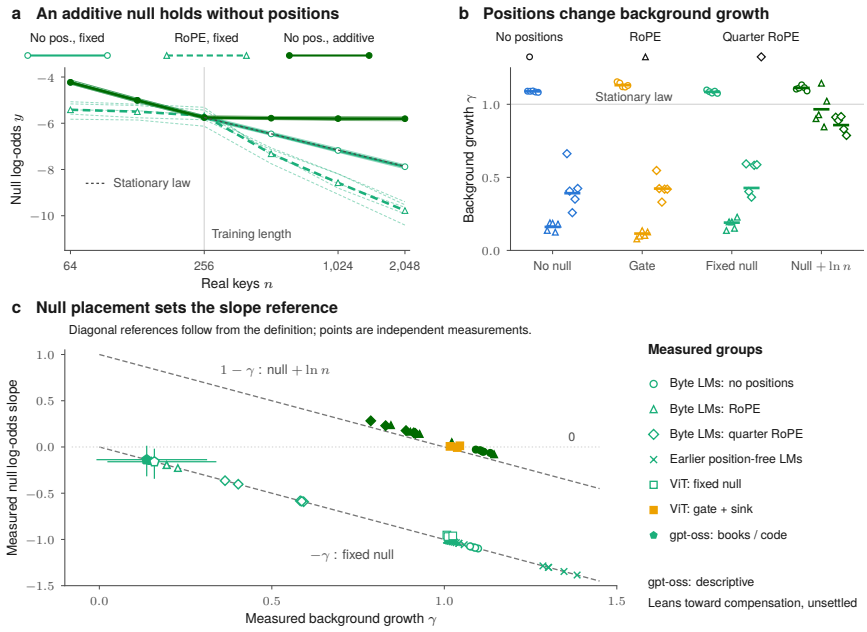


Figure 9.1: **Positions change the background's growth, and null placement sets how that growth affects abstention.** (a,b) 0.6M-parameter byte language models, 5 seeds per arm, trained at 256 tokens on one corpus. (a) Measured null log-odds $y = \log[p\ /(1-p\ )]$: thin curves are seed medians over heads, bold curves their mean. The additive null changes the fixed null by $\ln(n/N_{\text{train}})$ only beyond training, with no positions; the vertical rule marks the training length. The dashed stationary-law reference has slope -1.02, measured in the earlier position-free models trained at the same length, and is anchored to the fixed-null mean at training. (b) Background log-sum slope γ over 64–256 keys: each dot is a seed's median head slope; the short ticks are medians over all heads and seeds, as in the assessment. Shapes identify positions. Colors retain the null rule across panels. (c) Both coordinates are measured independently over matching key counts. The two diagonal references follow from the definition of log-odds for a fixed null logit $(-\gamma)$ and a null logit growing as $\ln n\ (1-\gamma)$; the horizontal reference is zero slope. They are not fitted trends or a theorem that gates are constant. Byte-LM circles/triangles/diamonds are the 5 seeds of each positional null arm over 64–256 keys (fixed null open, $+\ln n$ filled). Crosses are all 25 earlier position-free models (training lengths 64–1,024); each fits the saved counts $\{128, 512, 2048, N_{\text{train}}\}$, removing repeats, then takes the median head slope. Squares (fixed null open, gated filled) are 6 checkpoints (3 seeds per rule) of one ViT-S/16 size, untempered CLS rows over 224–896 px, beyond the 224 px training resolution; they use recorded medians over images, heads and all 12 layers on the first 1,000 images. For the ViT, the background sums patches only, while null odds include CLS among the real keys; the fixed-null points therefore need not lie on $-\gamma$. The two large pentagons are gpt-oss books (filled) and code (open) medians with marginal interquartile ranges over the same 700 sink-using dense heads, of the slope between query positions 256 and 4,096 in natural documents, which is shallower than the least-squares medians over every position of that range (Table 17.5): descriptive; the prespecified test on these heads was undecided by its rule. The byte-model contrasts are prespecified; the cross-model synthesis and ViT null slopes are descriptive. Causal interpretation is limited to the controlled positional contrasts.

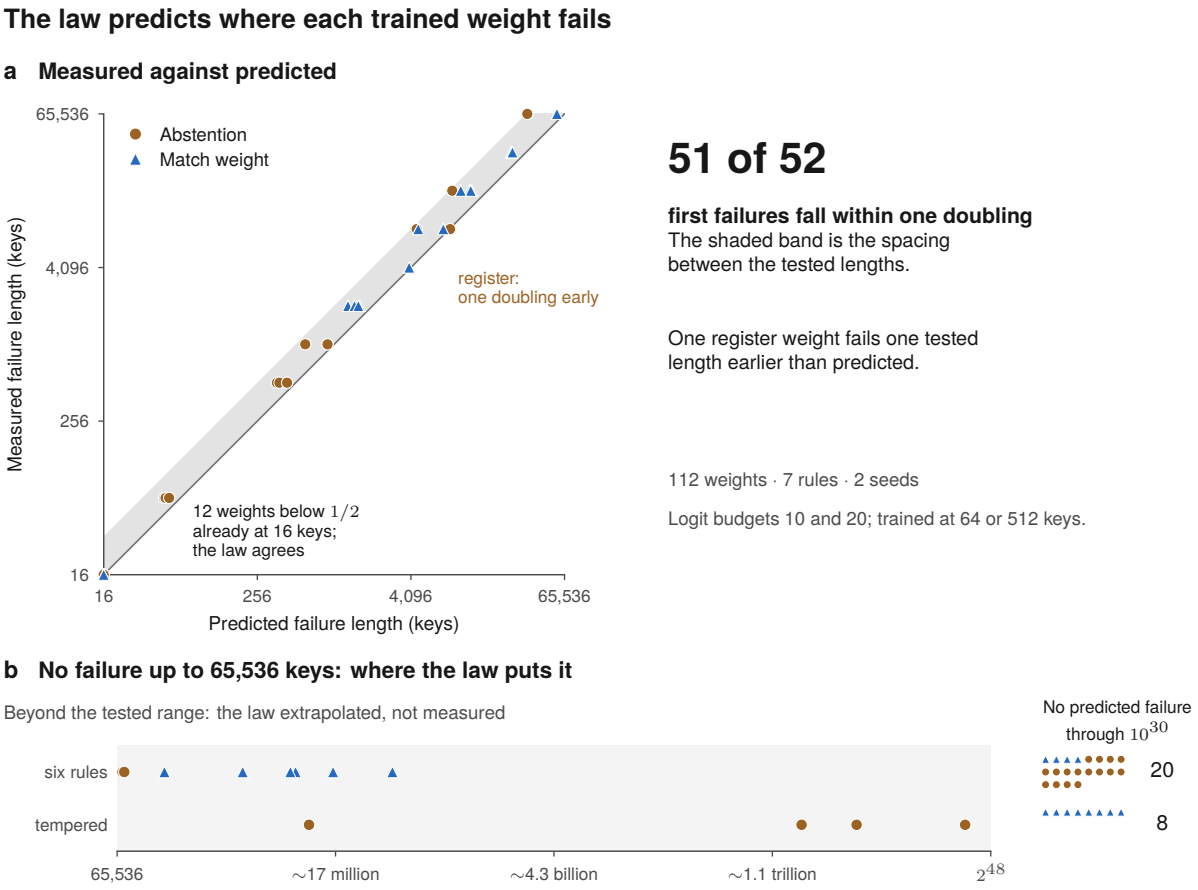


Figure 9.2: **The mean-field law predicts where each trained head's weights fail.** (a) For the six untempered rules, measured failure length against the prediction from each head's learned parameters, with nothing fitted: abstention on absent queries (brown circles) and match weight on present queries (blue triangles). Of the 52 weights that first fall below one half after the shortest tested length, 51 fall between the prediction and twice it (shaded), the resolution of the evaluated lengths. The remaining register weight fails one doubling before the first evaluated length at or above its prediction. Another 12 weights are already below one half at 16 keys, as the law predicts. (b) The 48 weights with no measured failure through 65,536 keys: the law predicts a later failure for 12 of the six untempered rules' weights (about $2^{16.2}$–$2^{26.1}$ keys) and all 8 tempered abstention weights (about $2^{23.0}$–$2^{47.1}$ keys). For the remaining 28, including all 8 tempered match weights, the law has no failure through 10^{30} keys; this is not an infinite-range guarantee. These are extrapolations of laws checked only to 65,536 keys, not measurements. The tempered prediction uses the numerically checked stationary power-mean law and does not certify the trained margins over the background. All 112 weights are represented: seven rules, logit budgets 10 and 20, training lengths 64 and 512, and 2 seeds, with 1,024 evaluation queries per length and seed. The shading is not a confidence interval; the quantities are mean attention weights, not task accuracies.

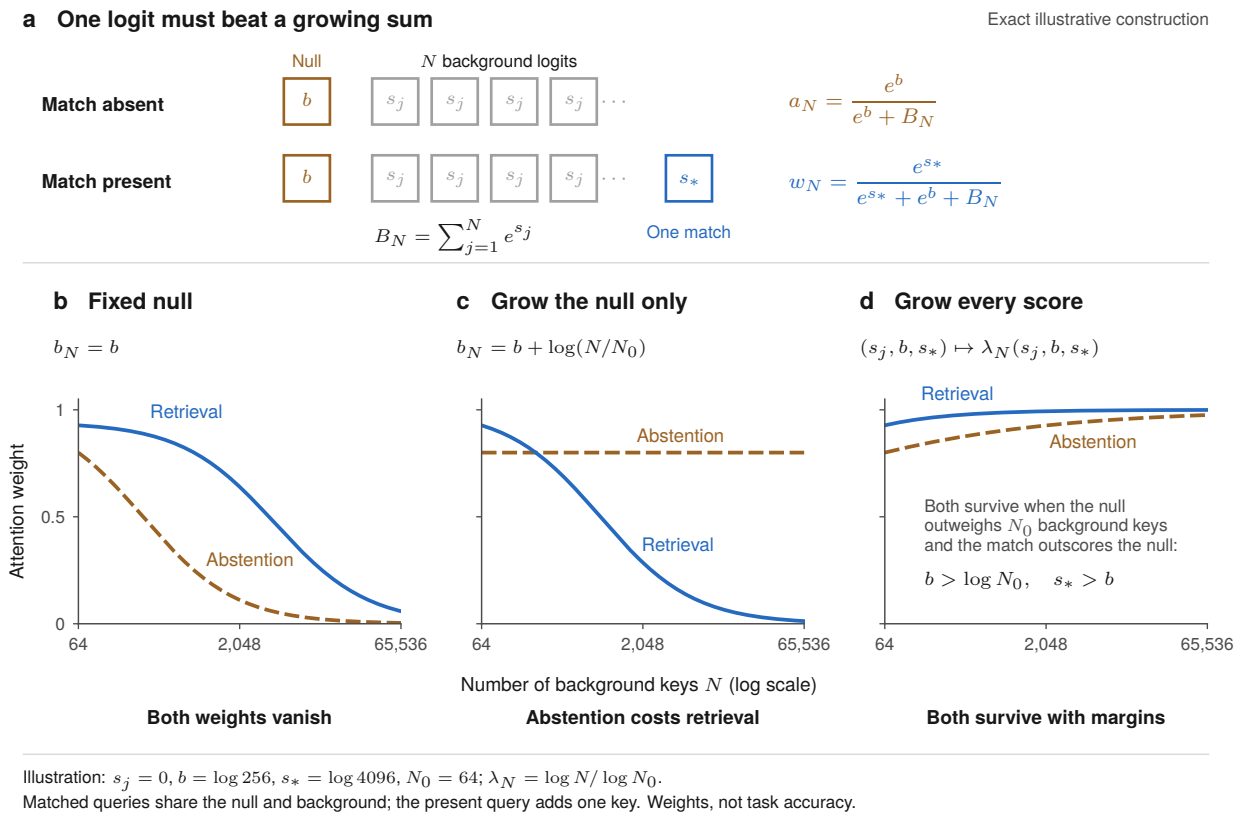


Figure 9.3: **Growing the null alone trades retrieval for abstention; growing every score can preserve both when the null and the match keep separated margins over the background. (a)** Matched absent and present queries share N background keys with scores s_j and a null score b; the present query adds one match with score s_*. Squares denote scores, not weights. The background sum is $B_N = \sum_j e^{s_j}$. Brown dashed curves show the absent query's null weight a_N; blue solid curves show the present query's match weight w_N. **(b)** Fixed scores eventually lose both weights. **(c)** Adding $\log(N/N_0)$ only to the null holds abstention constant while retrieval vanishes. **(d)** Multiplying every score, null included, by $\lambda_N = \log N / \log N_0$ preserves both in this exact illustrative construction: $s_j = 0$, $b = \log 256$, $s_* = \log 4096$, $N_0 = 64$. Its sufficient-margin factors are $N_0 e^{-b} = 1/4$, $N_0 e^{-s_*} = 1/64$ and $e^{b-s_*} = 1/16$, all below one. These are attention weights, not task accuracies; no measurements are drawn. Section 17.2 tests the three regimes in trained heads, where tempering's success is measured rather than guaranteed. The law covers rules that decide by summing over the keys; a gate that sees whether the match is present, or a head that reads only the largest score, lies outside it (Sections 9.1 and 9.4).

keys the onset has a closed form. Let the background keys be uniform on the unit sphere in $\mathbb{R}^D$, write F_D for the distribution function of the cosine between a fixed unit vector and a uniform one, and let the target's key have cosine $c \in (0,1)$ with the query. Then $N_{\text{ev}} \approx \ln 2/(1-F_D(c))$, and by Proposition 9.12(b)

$$-\ln(1-F_D(c)) = (D-1)\,C(c) + \ln\big(c\sqrt{2\pi(D-1)}\big) + O(1/D),$$

with $C(c) = -\frac{1}{2}\ln(1-c^2) = \frac{1}{2}\ln(1+\text{SNR})$ and $\text{SNR} = c^2/(1-c^2)$. The onset grows with the dimension at the capacity of a Gaussian channel at that signal-to-noise ratio (Shannon, 1959), not at its small-c form $c^2/2$. Random keys form a spherical code, and the evidence wall is the limit of decoding it.

The code limit has been measured. A registered test drew explicit random keys: 400 trials in each of 20 settings of c and D. In every setting the argmax's measured onset matched the exact law $F_D(c)^{N-1}$, with largest $|z|$ 2.80 against a family-wise threshold of 4.06 (Figure 9.6a). Its slope in D, after the prefactor $\ln(c\sqrt{2\pi(D-1)})$ is removed, was the capacity, not its leading order: 0.047, 0.107, 0.211 and 0.499 at $c = 0.30$, 0.44, 0.60 and 0.80, against $C(c) = 0.047$, 0.108, 0.223 and 0.511, where $c^2/2$ would give 0.045, 0.097, 0.180 and 0.320. The two rates are close at small c; the test named $c = 0.6$ and 0.8 in advance as the values that tell them apart. The wall is close at practical sizes. At $D = 128$ the strongest of a million random keys has median cosine 0.408 over 100 trials (exact median 0.410). Among a million keys, a target matched more weakly than that is outscored at least half the time by the best background key, and then no reader that ranks the scores can find it.

Soft routers spend a second budget. Fix the scores, with no ties at the target's.

Proposition 9.11 (Gain can be exchanged)**.**

(a) Select. *Draw M independent samples from a routing distribution that gives the target mass w and the items scoring above it total mass q, and keep the sample with the largest score. This finds the target with probability $(1-q)^M - (1-q-w)^M$. When $q = 0$ it succeeds at least half the time exactly when $w \geq w_M = 1 - 2^{-1/M}$.*

(b) Average. *A readout that needs a strict majority of the M samples on the target succeeds with probability $\Pr\{\text{Bin}(M,w) > M/2\}$. Its half-success threshold in w is never below $\frac{1}{2}$: exactly $\frac{1}{2}$ for odd M and higher for even M (0.614 at $M = 4$).*

(c) Trees. *If every internal node weights its children by their subtrees' partition functions, the leaf probabilities equal the flat softmax for any scores. A two-level product shortlist with $k \geq 2$ entries per half that contains the target in both halves, with target half-scores $c/2$ and other shortlisted half-scores at most u, puts at least half the mass on the target over this bounded-score class exactly when $\tau(c-2u) \geq 2\ln((1+\sqrt{2})(k-1))$.*

The proof is in Section A.10. A router that draws samples is a *walker*: a select walker keeps its best sample, and an averaging walker needs a majority of its samples on the target. In the mean-field model of this paragraph, N counts background items. Their sum is replaced by $Nm(\tau)$, with $m(\tau) = \mathbb{E}e^{\tau z}$ for a background score z and c the target's score. There the soft router's majority onset is $\ln N^* = \tau c - \ln m(\tau)$, its gain wall. Selection moves the wall to $\tau c - \ln m(\tau) + \ln(1/w_M - 1)$. Since $w_M \approx \ln 2/M$, each nat of $\ln M$ buys a nat of $\ln N$. Averaging buys nothing: by (b) each query still needs a majority of its mass on the target. By (c), a tree whose nodes weight their children by subtree partition functions is the flat softmax, so a tree's advantage lives in its hard stage. Even so, a soft final stage over a $k \times k$ shortlist needs $2\ln((1+\sqrt{2})(k-1))$ nats of separation:

3.96 at $k = 4$ and 7.18 at $k = 16$. Arbitrary learned parent scores do not reduce to the flat softmax, and Section A.10 gives an example. Two premises bound these statements. The threshold w_M, and with it the select wall, needs $q = 0$. When sampled distractors can outrank the target, $(1 - q)^M \to 0$ and more samples eventually hurt. The mean field is certified only where the background sum concentrates. For independent background scores with $m(2\tau) < \infty$, Chebyshev's inequality bounds the probability that its relative error exceeds u by $(m(2\tau)/m(\tau)^2 - 1)/(Nu^2)$. At large τ the sum is dominated by its few largest terms. For Gaussian scores $z \sim \mathcal{N}(0, 1/D)$ the mean-field onset $\tau c - \tau^2/(2D)$ peaks at $\tau = cD$; beyond that the background sum freezes, and the true onset keeps rising toward the evidence wall instead of falling (Lucibello & Mézard, 2024). That is an explanation within the Gaussian surrogate, not a finite-sample theorem for bounded keys. One part holds for any keys at every N: a majority reached at some τ is kept at every larger τ, so the true onset never falls as τ grows.

Constructed-key records. A product substrate stores keys as two 64-dimensional halves ($D = 128$) and queries them exactly ($c = 1$). A key's score is the mean of its two half-cosines, so the flat softmax factorizes over the halves, as in the proof of Proposition 9.11(c). The mean-field walls for softmax and for a select walker with M samples then become

$$\ln N^* = 2\ln\bigl(1 + (\sqrt{2} - 1)\, e^{\tau c/2 - \ln m(\tau/2,64)}\bigr), \qquad \ln N_M^* = 2\ln\bigl(1 + (w_M^{-1/2} - 1)\, e^{\tau c/2 - \ln m(\tau/2,64)}\bigr),$$

with $m(\cdot, 64)$ taken over one half. Softmax's wall sits there: measured at $\ln N = 7.90$, 12.36 and 16.67 at $\tau = 10$, 15 and 20, against 7.89, 12.37 and 16.69. Selecting walkers gain 1.066 and 1.070 nats of $\ln N$ per nat of $\ln M$ over $M = 16$ to 4,096 at $\tau = 5$ and 10, with every onset within 0.07 nats of $\ln N_M^*$. Averaging walkers do not move the wall outward, and with $M = 4$ they move it inward, since by Proposition 9.11(b) their half-success threshold at $M = 4$ exceeds one half. The registered prediction that averaging matches softmax at every M therefore fails (Figure 9.4a). A tree with a hard top-four shortlist per half shows no gain wall up to 2^{30} slots at $\tau \geq 10$. With a top-16 shortlist, its soft final stage still fails at $\ln N = 12.37$ at $\tau = 10$, consistent with Proposition 9.11(c). With weaker queries the machines' onsets approach the hard argmax's evidence wall, at $\ln N = 8.372$ for $c = 0.35$ and 18.006 for $c = 0.5$ (Figure 9.5). Select walkers with enough gain and samples reach it to within the grid's resolution. The registered convergence of the high-currency machines is only partly met: at $\tau = 120$ softmax stays 0.23 and 0.81 nats short, against a registered tolerance of 0.5. These tests fix c, so they lie outside the premises of Proposition 9.10(b), and how close a walker comes to the wall is observed here, not implied by that proposition. The accepted recovery supplies the full curves in Figure 9.4. Its saved selection slopes are 1.066 and 1.070 at $\tau = 5, 10$; the $\tau = 15$ slope is unscored because crossings are censored. The registered primary check, that every historical verdict keeps its status on the fine grid, fails: seven historical statuses change, all in the registered comparison of averaging walkers with softmax at $\tau = 15, 20$. Every check of the select walkers keeps its status, and no new fit or assessment is made here. For weak evidence, the recovered hard-argmax onsets are 8.372 and 18.006 nats at $c = 0.35, 0.5$ (Figure 9.5). At $c = 0.5$, softmax $\tau = 120$, selection $\tau = 20, M = 4096$ and the soft-final tree $\tau = 60, k = 4$ have saved onsets 17.198, 16.476 and 16.157. Their nearest measured points are $\ln N = 17.155, 16.462, 16.115$, with saved $(P_{\text{arg}}, S_{\text{arg}}, S_{\text{off}})$ triples $(0.582, 0.863, 0)$, $(0.650, 0.732, 0.079)$ and $(0.690, 0.734, 0)$, respectively. Here P_{arg} is target maximality, and $S_{\text{arg}}, S_{\text{off}}$ are success conditional on it and its complement: success is $P_{\text{arg}}S_{\text{arg}} + (1 - P_{\text{arg}})S_{\text{off}}$. Thus selection can succeed off the argmax, while failure even with a maximal target contributes to each displayed shortfall. These are saved readouts at measured points, not interpolation at a crossing or universal bounds. The historical convergence prediction remains partial. The recovery's overall assessment and execution gate remain failed; the registered exception admits publication as a rebaseline, not historical reproduction, with the bounded comparison/endpoint limitations in the source paper (Bruce, 2026e, Appendix C.2).

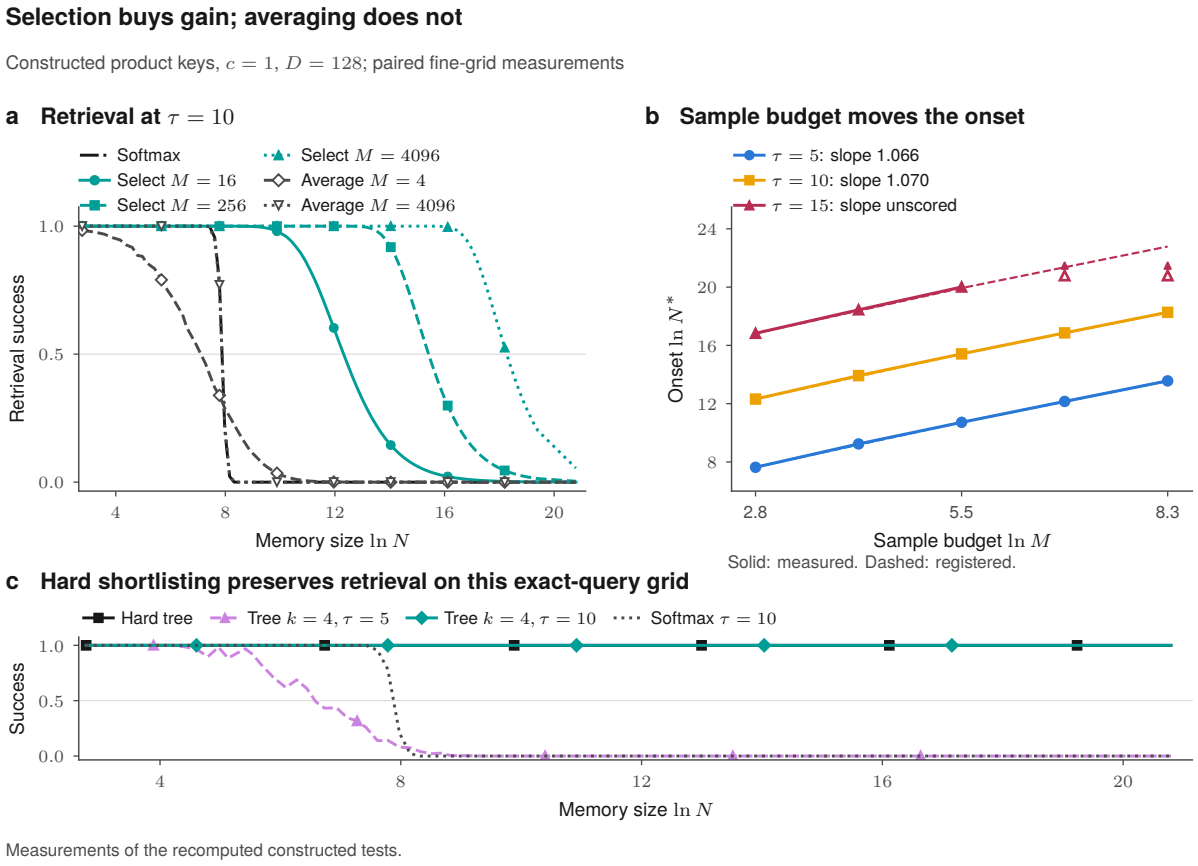


Figure 9.4: **Selection exchanges samples for retrieval capacity; averaging does not.** Recovered constructed product keys have $D = 128$ and exact queries ($c = 1$). (a) Retrieval success at $\tau = 10$: softmax, selection with $M = 16, 256, 4096$, and averaging with $M = 4, 4096$. (b) Saved first downward 50% crossings, interpolated in $\ln N$, versus $\ln M$ at $\tau = 5, 10, 15$; dashed lines are registered predictions. Open symbols and upward arrows mark no crossing by $N = 2^{30}$. Saved slopes are 1.066 and 1.070 at $\tau = 5, 10$; the $\tau = 15$ slope is unscored. (c) Hard tree, top-four-per-half soft trees at $\tau = 5, 10$, and softmax at $\tau = 10$. Softmax and soft-tree success require half the target mass; averaging requires a strict sample majority. Each point pools three seeds with 2,000 queries each; conditional walker probabilities are evaluated exactly. No seed intervals were retained. The overall assessment remains failed, and so does its primary check; publication is admitted as a rebaseline, with historical records not reproduced (Bruce, 2026e, Appendix C.2). These are synthetic tests.

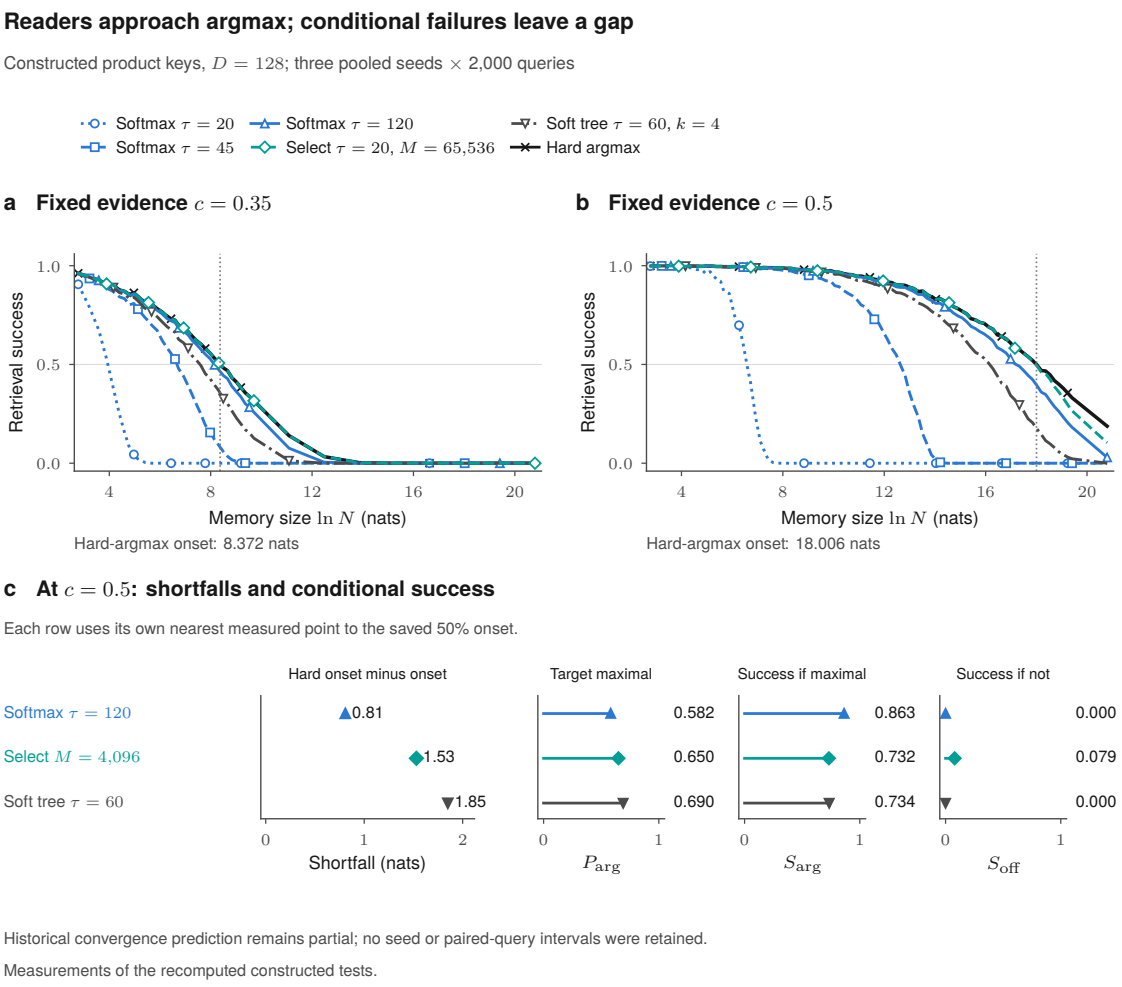


Figure 9.5: **Recovered readers approach the hard-argmax reference, with conditional failures explaining the remaining shortfall.** Constructed product keys have two 64-dimensional halves ($D = 128$). (a,b) All saved fine-grid points, merged with retained coarse points, at $c = 0.35, 0.5$: softmax $\tau = 20, 45, 120$; selection at $\tau = 20, M = 65{,}536$; a hard top-four shortlist per half with a soft final read at $\tau = 60$; and hard argmax. The vertical references are saved hard-argmax 50% onsets, 8.372 and 18.006 nats. Each curve pools three seeds with 2,000 queries each; softmax and soft-tree success require half the target mass, and walker probabilities are exact conditional on each query. (c) At $c = 0.5$, the three named readers use softmax $\tau = 120$, selection $\tau = 20, M = 4{,}096$, and soft tree $\tau = 60, k = 4$. Left: signed difference between the saved hard and machine onsets. Right: P_{arg} is the probability the target is maximal, and $S_{\text{arg}}, S_{\text{off}}$ are success conditional on that event and its complement. These probabilities are saved at the nearest measured point to each onset, not interpolated; conditional counts are 3494/2506, 3898/2102 and 4142/1858, respectively. No seed or paired-query intervals were retained. Pooled excesses are not clamped; these fixed-c tests are outside the premises of Proposition 9.10(b). The historical convergence prediction remains partial. Execution remains failed; publication is admitted as a rebaseline, with historical records not reproduced (Bruce, 2026e, Appendix C.2).

The two budgets are the two limits of decoding. The code limit is the evidence wall: it depends on the keys, through D and c, and under part (b) of Proposition 9.10 no read moves it. The decoder limit belongs to the read. A soft read whose logits lie in a window of width R can put a majority on one of N items only if $\ln(N-1) \le R$ (Proposition 9.12(c)). Raising τ, or growing it with $\ln N$, widens the window and lifts the decoder limit toward the code limit; in the registered test it never passed it. A read that takes the maximum over all items has no decoder limit. The registered test of random keys measured both. Raising τ from 2.5 to 80 moved softmax's onset from 3.4–10.0 nats below the evidence wall to 0.02–0.81 nats below it, never above. Its onset followed the mean-field wall $\tau c - \ln m(\tau, D)$, with the moment taken in dimension D, in all 57 registered comparisons inside the mean-field regime, with largest deviation 0.40 nats against a band of $\ln 2$.

A second registered stage trained lookup heads on noisy-match geometries, with D varied: four rules, logit budgets $\tau_{\max} \in \{10, 20, 40\}$ and three seeds, 360 cells in all. It reads accuracy and weight separately. On accuracy, no rule outlasts the code limit by more than one grid step, and a head that reads the maximum (a key check) and a tempered sum reach the limit at every D, to within one grid step. Their slopes in D, with the prefactor removed as above, are 0.215 and 0.216 against $C = 0.223$ at match cosine 0.6, and 0.498 and 0.507 against 0.511 at 0.8, where $c^2/2$ would give 0.18 and 0.32. A sum at a fixed scale falls behind as D grows. At $\tau_{\max} = 10$ its accuracy onset sits 0.12–0.16 nats below the code limit at the smallest D and 1.06–1.44 nats below it at the largest. These gaps are seed means of crossings interpolated on a grid of 0.69-nat steps, reported but not scored; the scored test found the sum at least one grid step short at the largest D in every seed. On weight, the decoder limit binds a trained sum. Of the 220 scored cells, the 11 in the concentration regime fall on the trained law, two-sided within a grid step. The rest fail no earlier than one grid step before it, as Jensen's inequality allows: outside that regime the law bounds the failure on one side only. On abstention, in a post hoc reading of the saved cells, tempering does not rescue these heads. Its factor is one at the training count of 64 keys, so in all 90 pairs the tempered head has its fixed-sink partner's weights, and its absent-query abstention falls with N as the fixed sink's does: from a median of 0.59 at the training count to at most 0.005 at $N = 65{,}536$, against at most 0.013 for the fixed sink. With exact queries the tempered sink keeps abstention over the tested range (Section 8.4); with these noisy matches it does not.

9.4 Sums need a $\log N$ schedule; maxima need none

Every rule in Sections 9.1 and 9.2 decides by comparing one logit with a sum over the N keys: softmax, sinks, registers, gated heads and sigmoid weights. A head that decides by reading a maximum, abstaining unless its best score clears a threshold, needs no schedule against backgrounds bounded below its threshold, and otherwise only a threshold that tracks the background's extreme value; it fails only when the background maximum reaches the threshold (Section 8.4; Bruce, 2026e). In the Gaussian surrogate, a softmax whose scale grows past the point where the background sum is carried by its largest terms reads the maximum in the same way. The harness included a head that reads a maximum: with YaRN it was the only rule whose null and match weights both changed by less than half a nat of log-odds out to eight times the training length, on means over 3 seeds though not in every seed (five rules kept the match weight alone better), and without YaRN it failed as the matching key fell far down the ranking and the background maximum grew with length (Section 17.4). The usual answer that softmax can concentrate almost all of its weight on one position, which exact copying and retrieval need (Section 0.2.2), is quantified by this chapter: the margin must grow like $\log N$ (Section 9.1), no router beats the argmax on the same scores, under part (b) of Proposition 9.10 (Section 9.3), and sharpness trades against abstention (Sections 9.1

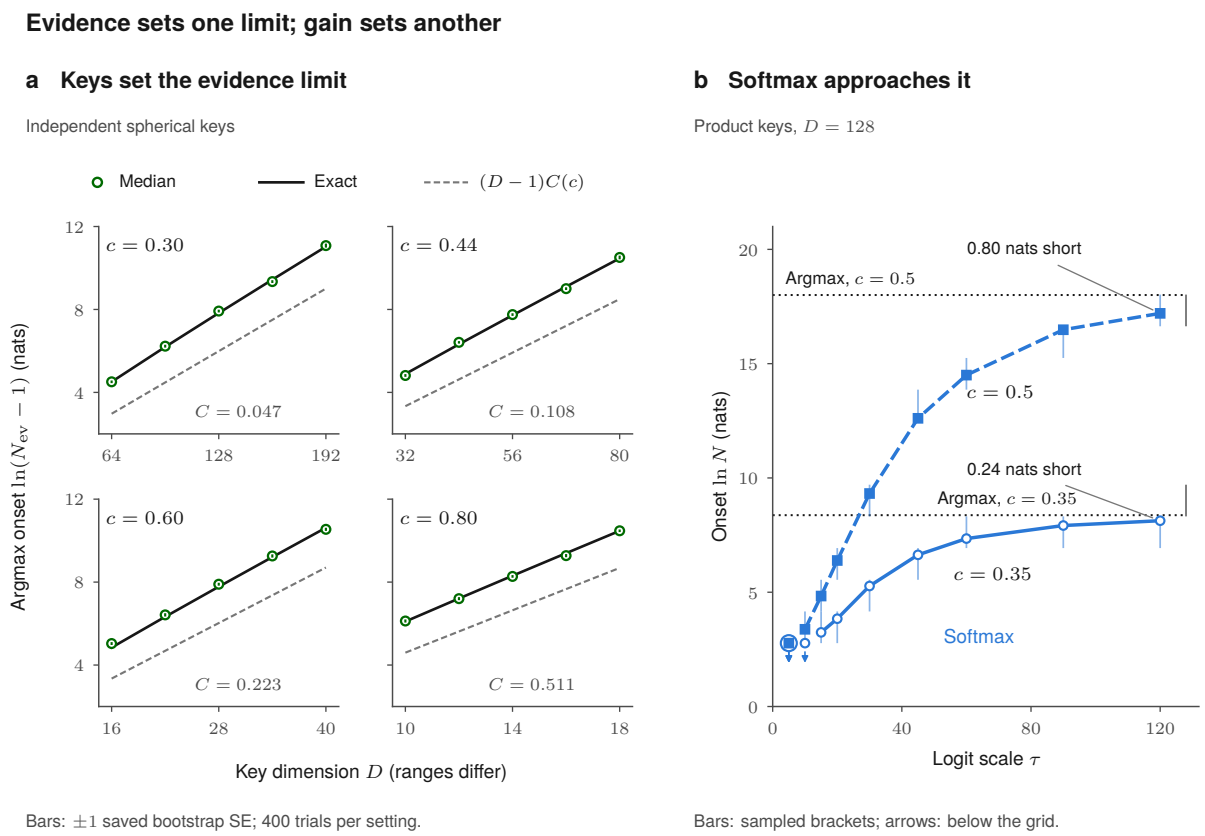


Figure 9.6: **Keys set the evidence limit, while increasing softmax gain approaches it on these synthetic substrates.** (a) Independent spherical keys, the registered test of Section 9.3: all 20 settings, with 400 trials each. Open circles show the log median first background exceedance T_{ev}; bars are ±1 saved bootstrap standard error (2,000 resamples, seed 29), often smaller than the symbol. This background-count convention is $\ln(N_{\text{ev}} - 1)$, matching the exact half-success law $N_{\text{ev}} - 1 = \ln 2/[-\ln F_D(c)]$ (solid). Dashed lines show the capacity term $(D-1)C(c)$, where $C(c) = -\frac{1}{2}\ln(1-c^2)$; finite-size prefactors separate it from the exact law. Reference lines join values at the tested dimensions; the four facets share their vertical scale but use different dimension ranges. (b) The separate product-key test ($D = 128$, two 64-dimensional halves): softmax's first downward 50% success crossing, interpolated in $\ln N$, against gain τ. Success requires at least half the target mass. Dotted lines are the measured argmax crossings. Blue bars and the rightmost grey bars span the sampled sizes bracketing softmax and argmax crossings, respectively; they are grid-resolution brackets, not confidence intervals. Downward arrows mark failure already at $N = 16$: both cosines at $\tau = 5$ (nested symbols), and $c = 0.35$ at $\tau = 10$. Each success value pools three seeds × 2,000 queries; seed-level spread was not retained. At $\tau = 120$, softmax remains 0.23 and 0.81 nats below the respective walls.

and 9.2).

A schedule repairs the read; it adds no evidence to the keys. For a random target score, averaging part (c) of Proposition 9.10 over it gives the argmax curve, and under part (b) no rule beats that curve. So, under (b), samples, trees, schedules and more gain can bring a router up to the argmax curve but never past it, and only better keys move it.

The budget of Theorem 9.1 is a limit of the read, not of the keys. A soft rule whose log-odds between two keys are capped at 2τ cannot certify more than about $e^{2\tau}$ alternatives, whatever the keys are; tempering raises the cap, and a head that reads a maximum has none. The other limit, the code limit of Section 9.3, belongs to the keys. When the match's score is noisy and its log-likelihood ratio is linear in the score, the posterior over 'no match' and each key is a softmax at the true noise level whose null logit is $\log N$ plus a constant, and thresholding it is the optimal decision with an abstention option (Forney, 1968). With noise-free matches there is no finite true temperature: the best read that never ranks a lower score above a higher one is the maximum, which tempering approaches from below, as random-energy analyses of softmax retrieval also find (Lucibello & Mézard, 2024). Either way, what fails at long context is a fixed temperature and a fixed null, not the summing.

Figure 9.7 draws the law of sums on exact constructions, with two of its neighbours: a sum pays for the size of its population while a maximum does not, a head that routes over its top k keys and gates on its largest score separates routing from abstention with no length schedule built in, and order carried in soft weights has a positional price, which Section 10.7 develops.

9.5 Routing information is conserved

Whatever the machine, naming one of N items costs about $\ln N$ nats; what differs is the currency. A store holds N items, item i with key k_i and value x_i. A query names one item, the target J, or none. All quantities are in nats, and $h(\varepsilon) = -\varepsilon \ln \varepsilon - (1-\varepsilon)\ln(1-\varepsilon)$ is the binary entropy. Each part of the following proposition prices one task within its own class of machines: what must be paid, and in which currency. The parts form a family of lower bounds, each holding within its own class, and they set no exchange rate between currencies.

Proposition 9.12 (Conservation of routing information)**.**

(a) Identification. *Let J be uniform on $\{1, \dots, N\}$ and let a rule, randomized or not, name $\hat{J}$ from an observation Y. If $\Pr(\hat{J} \neq J) \leq \varepsilon$, then $I(J;Y) \geq (1-\varepsilon)\ln N - \ln 2$.*

(b) The key channel. *Let the $N-1$ background keys be independent and uniform on the unit sphere in $\mathbb{R}^d$, independent of the query q, and let the target's key have cosine $c \in (0,1)$ with q. Consider readers that rank keys by their cosine with q: an argmax, a top-k stage, a gate on the largest score, or a soft router whose weights increase with the score. Such a reader's final selection is the target, or it puts a strict majority on the target, only if the target is the argmax. That has probability $F_d(c)^{N-1}$, where F_d is the distribution function of the cosine, and the probability is one half at $N_{\mathrm{ev}} = 1 + \ln 2/(-\ln F_d(c))$. With $C(c) = -\frac{1}{2}\ln(1-c^2) = \frac{1}{2}\ln(1+\mathrm{SNR})$ and $\mathrm{SNR} = c^2/(1-c^2)$,*

$$-\ln\bigl(1 - F_d(c)\bigr) = (d-1)\,C(c) + \ln\bigl(c\sqrt{2\pi(d-1)}\bigr) + \frac{c^{-2} - \frac{3}{4}}{d-1} + O(d^{-2}).$$

(c) Expression. *Let a soft router give the target the weight $w_t = e^{s_t}/(z + \sum_{j=1}^{N} e^{s_j})$, where $z \geq 0$ is an optional null weight and all N logits lie in a window of width R. If $w_t \geq w^*$, then*

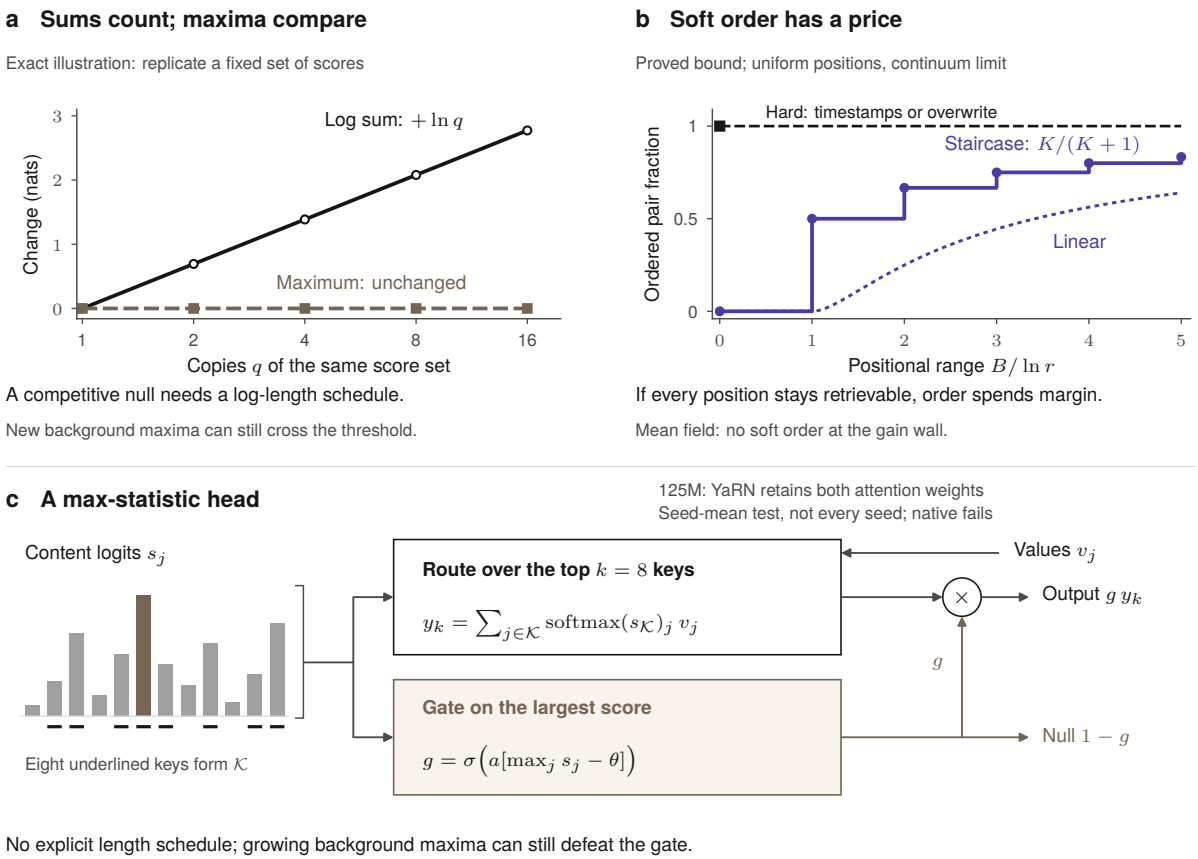


Figure 9.7: **Sums pay for population size; a max-statistic head separates routing from abstention, while soft order has a positional price.** Exact illustrative constructions, not experimental measurements. **(a)** Replicating a fixed score set q times adds $\ln q$ to $\ln Z$, $Z = \sum_j e^{s_j}$, and leaves $\max_j s_j$ unchanged. A competitive null's logit must grow by $\ln q$ to keep its mass fixed. New extreme background scores can still cross a maximum-based test's threshold. **(b)** For two identical writes, ordered means that the later has at least $r > 1$ times the earlier weight. With positional range B, $K = \lfloor B/\ln r \rfloor$, the optimal equal-block staircase orders $K/(K+1)$ of pairs in the uniform-position continuum limit; a linear bias orders $(1 - \ln r/B)_+^2$. Hard timestamps or overwrite order every pair without spending positional-logit margin. The gain-wall statement additionally assumes mean field and retrieval of a single target at every position (Proposition 10.6). **(c)** Illustrative logits feed the specified top-k route ($k = 8$, underlined keys) and a largest-score gate. Values v_j enter separately; σ is logistic and a, θ are learned per head. The output has mass g and the null has mass $1 - g$. No explicit length schedule is built in. In trained 125M language models, the registered secondary retention test for null and match weights passed with YaRN on mean per-seed log-odds changes over three seeds, but not in every seed. Native extrapolation failed as the background maximum grew with length. These are attention weights from fallback heads, not behavioural abstention or retrieval, and give no guarantee beyond the tested lengths (Section 17.4).

Table 9.1: Who pays which ledger of Proposition 9.12. S is the state size in nats and v the nats per stored value. Every machine that ranks keys stops at the code limit at the latest.

Machine	How it pays	Where it stops
Softmax attention	Memory in full, with a cache that grows with N; expression with its logit scale or a $\ln N$ schedule; recency from positional range	Its decoder limit unless the scale grows; then the code limit
Fixed-state recurrence	Direction and recency by construction	The memory wall, once $N(v - h(\varepsilon) - \varepsilon \ln(e^v - 1)) > S$, unless the question comes first
Max-statistic head	Nothing for expression: it selects	The code limit
Timestamps	Order, with $N \ln N \geq \ln N!$ nats of storage	—
Overwrite	Never pays for order: it answers latest-value queries only	It loses the history

$N - 1 \leq e^R(1 - w^*)/w^*$. *A majority* ($w^* = \frac{1}{2}$) *therefore needs* $\ln(N - 1) \leq R$. *A reader of type (b) that selects pays nothing here.*

(d) Memory. *A one-pass machine reads the store into a state M that takes at most e^S values, possibly using randomness. The query arrives afterwards, and the answer is computed from M and the query alone. Let the keys be fixed and the values independent and uniform on a set of e^v elements. If each item's value is returned correctly with probability at least $1 - \varepsilon$, where $\varepsilon \leq 1 - e^{-v}$ (no worse than guessing), then*

$$S \;\geq\; N\big(v - h(\varepsilon) - \varepsilon \ln(e^v - 1)\big) \;\geq\; N\big((1 - \varepsilon)v - \ln 2\big).$$

For membership, let the stored keys be an N-subset of a universe of U keys, let the machine be deterministic, and let it report every stored key present and at most a fraction $\varepsilon \in (0, 1]$ of the absent keys. Then $S \geq \ln \binom{U}{N} - \ln \binom{\lfloor N + \varepsilon(U-N) \rfloor}{N}$, which tends to $N \ln(1/\varepsilon)$ as $U \to \infty$ at fixed N.

(e) Order. *A one-pass machine reads N distinct, known writes, arriving in a uniformly random order, into a state of at most e^S values. It then answers "which of i and j came first?" from the state alone. If all its answers are correct together with probability at least $1 - \varepsilon$, then $S \geq (1 - \varepsilon) \ln N! - \ln 2$. A soft router that carries order in its weights is priced by Proposition 10.6.*

The proof is in Section A.10. The five parts are five ledgers. Naming the target with error at most ε takes at least $(1 - \varepsilon) \ln N - \ln 2$ nats about it (Fano's inequality). Readers that rank keys succeed only through the argmax, at the code limit of Section 9.3, whose dimension D is the d here. A soft majority needs a logit range of at least $\ln(N - 1)$. A one-pass state that answers after reading N items of v nats each, returning each value with error at most $\varepsilon \leq 1 - e^{-v}$ (no worse than guessing), needs $N(v - h(\varepsilon) - \varepsilon \ln(e^v - 1))$ nats, where h is the binary entropy, and membership with no false negatives and a false-positive rate ε needs about $N \ln(1/\varepsilon)$. A recoverable order of N writes costs $(1 - \varepsilon) \ln N! - \ln 2$ nats; a soft router pays for order in positional range instead.

Table 9.1 shows who pays which ledger, and Remark 9.13 gives the reasons.

Remark 9.13 (Who pays which ledger).

- *Softmax attention* pays memory in full, with a cache that grows with N. It pays expression with its logit scale or a $\ln N$ schedule, and buys recency from the positional range of Proposition 10.6.

- *A recurrent machine with a fixed state* gets direction and recency by construction. At an error $\varepsilon < 1 - e^{-v}$ it cannot pay memory once N passes $S/(v - h(\varepsilon) - \varepsilon \ln(e^v - 1))$, unless the question comes first, which takes it outside part (d) of Proposition 9.12: its state then needs to hold only the answer.
- *A max-statistic head* pays nothing for expression and stops at the key channel.
- *Timestamps* pay for order with $N \ln N \geq \ln N!$ nats of storage. *Overwrite* answers only latest-value queries, so it never pays part (e) of Proposition 9.12 and loses the history instead.

Proposition 10.6(d) says a hard machine spends none of the router's margin on order; part (e) of Proposition 9.12 says what it spends instead.

No fixed allotment of dimension, logit range, state or positional range survives a growing N. Part (v) below concerns order carried in soft weights: two identical writes are ordered with ratio $r > 1$ when the later one receives at least r times the weight of the earlier one, and $t(N, K+1)$ is the number of edges of the Turán graph, the complete $(K+1)$-partite graph on N vertices with parts as equal as possible (Proposition 10.6).

Corollary 9.14 (A wall for constant resources)**.** *Fix the dimension d and match cosine $c < 1$, the logit window R, the state size S and the positional range B. Each of the following holds once N is large enough.*

(i) A reader of type (b) of Proposition 9.12 succeeds with probability below one half (once $N > N_{\mathrm{ev}}$).

(ii) A soft router puts a majority on no target (once $N - 1 > e^R$).

(iii) A one-pass machine cannot return an arbitrary stored value with error at most $\varepsilon < 1 - e^{-v}$ (once $N(v - h(\varepsilon) - \varepsilon \ln(e^v - 1)) > S$).

(iv) It cannot keep a recoverable order (once $(1 - \varepsilon) \ln N! - \ln 2 > S$).

(v) A soft router cannot order a fraction $\phi > K/(K+1)$ of the pairs, with $K = \lfloor B / \ln r \rfloor$ (once $t(N, K+1) < \phi \binom{N}{2}$).

Each wall moves only when its own ledger is paid:

- *dimension or match quality for (i);*
- *logit range or a selecting read for (ii);*
- *state for (iii) and (iv);*
- *positional range for (v).*

Leaving the class also moves it: reading the question before the store, or reading the store twice, takes a machine outside (iii).

The proof is in Section A.10.

Remark 9.15 (Exchange rates, measured). Proposition 9.12 fixes the price; three registered tests measure how efficiently machines pay it (Figure 9.8); a later amendment to the memory test only tightened its assessment.

- *Memory.* A linear-attention (outer-product) state of d^2 numbers, written with random unit keys in $\mathbb{R}^d$ and independent values uniform on $\{-1, +1\}^d$, and read by signs, recalls 0.315–0.328 nat per number, close to $1/\pi$, once it holds $16d$ items or more. A least-squares state, the delta rule's fixed point, recalls every stored bit exactly up to $N = d$, which is d^2 bits at $N = d$, and then

falls to the same rate. The online delta rule with a unit update rate peaks at 0.272–0.275 nat per number at $N = d$ to $2d$, settles at 0.264–0.267 by $N = 64d$, and recalls what is recent: the signs of its newest d items are read with accuracy 0.80, those of its oldest eighth at chance.

- *Fuzzy addressing.* Identifying stored binary codes from noisy queries follows the exact binary-code law in all 20 settings (largest $|z|$ 2.44). With each query bit flipped independently with a fixed probability $0 \le p < 1/2$, its cost per stored key tends to $\ln N/(\ln 2 - h(p))$ bits, set by the capacity $\ln 2 - h(p)$ of the binary symmetric channel. The registered slope test failed at two of its four flip rates: against a 10% tolerance, the measured slope ran 14.4% and 10.4% above capacity at $p = 0.05$ and 0.1. The registration itself had noted that the exact law's own slope runs 0–6% above capacity at these code lengths. Post hoc and unscored, the measured slopes sit 2.5, 1.0 and 0.7 standard errors above the law's own at $p = 0.05$, 0.1 and 0.2: within the sampling noise the tolerance did not budget at the two larger flip rates, and at its edge at the smallest. The law itself converges to capacity as the code length grows.
- *Addressing positions.* For a unary feature, the largest logit range needed to put a majority on a target, among the positions checked, grows like n^2 (fitted exponent 2.01, about $0.78(n-1)^2$). A rotary code needs 5.7–25.6 nats up to $n = 4{,}096$ with $2\log_2 n$ dimensions, within 3.1 times the floor $\ln(n-1)$. Recency carried in soft weights is unary whatever the code. With $K \ge 2$ an optimized rotary term orders 72–89% of the pairs that the staircase of Proposition 10.6 orders, at $K = 1$ almost none (at most 0.1%), and never more; a linear bias leaves 2.16–2.29 times as many pairs unordered as the staircase.
- *Each machine at its own wall.* On one recall task, the argmax fails at the code limit, with slope 0.526 in d after subtracting the logarithmic prefactor, against $C = 0.511$. A one-slot machine that reads the question first decides exactly as the argmax, by construction; what it saves is the state, one slot in place of the whole cache. Softmax at a fixed scale fails at its decoder limit, which moves 1.4–2.5 nats per doubling of τ. An outer-product state that reads the question last fails at its memory wall, near 30–57 items and 34–713 times below the code limit; the wall grows about linearly with d. A wider value code moves only that wall: quadrupling it gains 0.21–0.24 nats, and the other machines' onsets stay put.

9.6 Retrieval's budget on length and resolution

Retrieval is the better-known half of the law. Scaling the logits with $\log N$ keeps attention sharp on long inputs (Chiang & Cholak, 2022; Su, 2021), and the same budget appears on the resolution axis of vision models (Wang et al., 2024; Leroy et al., 2024; Gao et al., 2024). On the length axis, a discrete retrieval task cannot test the law: none of the 48 models trained on it under two protocols learned it, in-length retrieval was at most 2.1%, and every recorded logit gap was negative. The test therefore uses a continuous one-hop retrieval task with a learned decoder. Fifteen models were trained at $N_0 = 32$, 64 and 128 keys, five seeds each, and evaluated on a doubling grid of lengths. Retrieval is the fraction of queries whose decoded output is closest in cosine to the partner's content, and its onset is the first length at which that fraction falls below 0.5. The native onsets are 8,192, 32,768 and 131,072 keys, the same in every seed at each training length. They grow as $e^{2.14\,g_0}$ in the trained logit gap g_0, which ranges from 4.18 to 5.54 (base-2 slope 3.08, descriptive 95% bootstrap interval [3.03, 3.15] over the fifteen models; Figure 9.9a). These are accuracy onsets of a learned decoder. The records hold no attention weights, so no weight onset is claimed, and the figure's $k = 1$ line is only a worst-case comparator. Multiplying the logits by $\log N/\log N_0$ at evaluation keeps retrieval at 0.9998 or above to 131,072 keys in all fifteen models. The native readout there

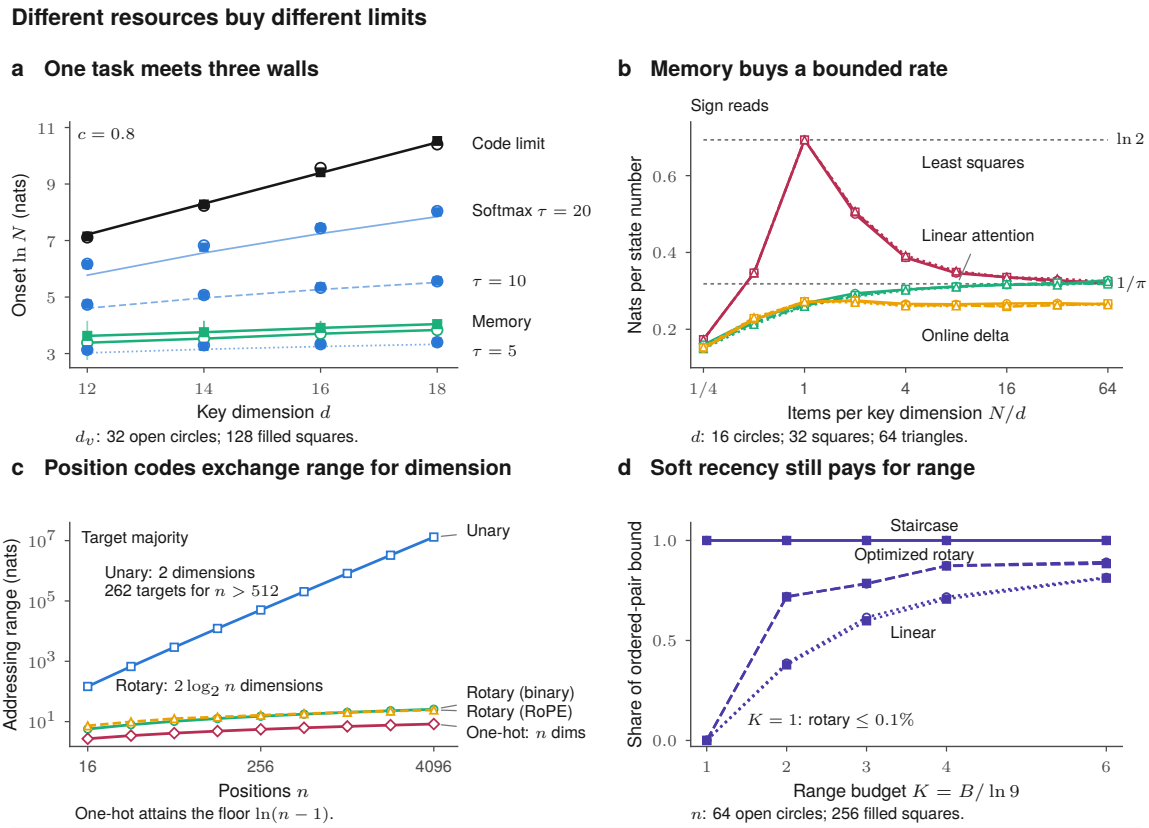


Figure 9.8: **Different resources buy different limits in three separate synthetic tests.** (a) The test of each machine at its own wall, all eight settings: cache argmax (black), softmax (blue) and a question-last outer-product state (green), at match cosine $c = 0.8$ and 16 value classes. Open circles and filled squares give value widths $d_v = 32$ and 128; widening values moves only the memory wall. The question-first one-slot maximum equals the cache argmax in every trial, by construction. Argmax and softmax each use 400 trials per setting. Their saved onsets are $\ell = \ln \text{median}\, T$ for the first background index T; the common total-count axis plots $\ln(1 + e^{\ell})$, and bar ends are $\ln(1 + e^{\ell \pm s})$ for the saved bootstrap standard error s (2,000 resamples, seed 59). These are transformed intervals, not new standard errors of $\ln N$. The black line is the saved total-count code law; pale blue lines are the saved mean-field decoder references, reported rather than scored and least reliable at $\tau = 20$. The memory state uses its own 4,000 trials: its half-success crossing is interpolated in $\ln N$, with bars spanning the sampled sizes, not confidence intervals. (b) The memory test, all 81 memory cells: age-binned sign-read information divided by d^2, a lower bound on recalled information, with 8–16,384 trials per cell and no saved trial intervals. The three dimensions are separate curves. Linear attention reaches 0.315–0.328 from $N = 16d$; online delta peaks at 0.272–0.275 and settles at 0.264–0.267 by $N = 64d$. (c) The positional-addressing test, all 36 cells. Unary shows the maximum required range over checked targets: every target through $n = 512$; 262 targets each at $n = 1{,}024$, 2,048 and 4,096 (every $(n/256)$-th zero-based target, plus $0, 1, 2, n-3, n-2, n-1$ and $\lfloor (n-1)/2 \rfloor$). The latter are sampled maxima, not all-target certificates. Rotary solutions use shared amplitudes and are constrained and checked at every target; the one-hot code attains the exact floor $\ln(n-1)$. (d) Its separate 30-cell recency test: recorded ordered-pair counts divided by the finite-n bound of Proposition 10.6, at ratio $r = 9$ and range $K \ln 9$. Rotary points are the best saved solutions from 32 restarts, not certified optima: 72–89% of the bound for $K \geq 2$, at most 0.1% at $K = 1$. Lines join tested settings. The footer preserves the registered slope failure of the fuzzy-addressing test; no slope is refitted here.

has fallen to at most 5.0×10^{-5}, to 0.0007–0.0011 and to 0.33–0.46 for the three training lengths. No rescaled onset was observed, so how far the rescale holds beyond 131,072 is unknown. The same fifteen models were first evaluated only to 16,384 keys; that run is an exact prefix of this one, not a replication. A third arm built the scale into the forward pass, but its factor is one at N_0, the only training length, so its fifteen models have the same weights as these, and its readout, 0.9999 or above to 16,384 keys, is the evaluation rescale's (Figure 9.9b). It is not evidence about training with the scale.

On the resolution axis, where N is the number of tokens and grows with the resolution, vision transformers trained from scratch at one resolution, five seeds for each dataset, lose on average 43.5 and 29.6 percentage points of accuracy at the highest test resolution on CIFAR-100 and Tiny-ImageNet. Their higher test resolutions are the test images upsampled bicubically, with the position embeddings interpolated, so they add tokens but no image content. Multiplying the logits by $\log N / \log N_0$ without retraining recovers 0.65 [0.37, 0.89] and 0.83 [0.66, 1.00] points on average across test resolutions above the training resolution (95% intervals over seed–resolution pairs), 3–5% of the deficit. Building the budget into multi-resolution training gains 6.3 and 6.4 points on average at the top resolution over the same training without it (five seeds a side). Pretrained ViT-T, ViT-S and ViT-B, evaluated once on a fixed subset of 5,000 ImageNet validation images, lose less between 224 and 384 pixels (4.56, 2.04 and 0.80 points). There the training-free rescale recovers 3.02, 1.95 and 0.12 points on average across test resolutions above 224 pixels and stays within 0.26 points of the best constant chosen for each resolution. These runs normalized their inputs with ImageNet's channel statistics, whereas the checkpoints expect a mean and standard deviation of 0.5, which depresses their accuracy, at 224 pixels included; the drops, gains and multipliers quoted for them carry that caveat. The optimal logit multiplier c^* rises with $\log N$ (Figure 9.9c). Over the 130 model-resolution pairs at which models degrade gracefully, its slope in $\log N / \log N_0$ is 0.77 [0.59, 0.95] (a 95% interval over these pairs, not over models). Pooled over these pairs, $\log N$ fits better than a form linear in N (R^2 of 0.39 against 0.27). But when fitted model by model, it fits better in only 10 of the 23 models. The slope depends on the recipe: 1.05–1.13 for the pretrained ViTs, 1.40–1.42 with the budget trained in, and 0.34–0.42 for plain multi-resolution training, the last two as five-seed means on each dataset. All ten models trained from scratch at one resolution collapse at three of their six test resolutions; these models are excluded from the fit, and the registered probe of the collapse regime's boundary was inconclusive.

Two predictions did not hold. The best fitted constant temperature matched or beat the $\log N$ rule with a fitted exponent, $(\log N / \log N_0)^p$, on extrapolated accuracy. Each was fitted at the first test resolution above the training resolution and scored at the higher ones: the fitted $\log N$ rule minus the constant was -0.44 points on average over the 10 models trained at one resolution (95% interval $[-0.83, -0.04]$). As registered, that makes the budget on this axis a tunable temperature with no privileged shape. The unfitted rule, $p = 1$, matched the fitted constant, with a mean difference of 0.01 points. And the budget did not carry over to the minimal block of Section 16.6: in models trained at one resolution, where its factor is one during training, switching it on changed top-resolution accuracy there by -0.28 and -0.21 points, against $+0.18$ and $+0.44$ for the ViT (means over five seeds), so this test measured the evaluation-time rescale. Two further tests were inconclusive. The heads of ViT-Tiny were too uniform for the per-head test (normalized entropy range 0.050, below its gate of 0.1), and no run reached the collapse-prone regime that the logit-magnitude test required.

Without a null, Theorem 9.1 reduces to retrieval alone: the match must outweigh the background, $e^{s_*} \geq W_* B_N$, so against background scores of a fixed level the margin must grow like $\log N$. A head that may circulate can route at that logarithmic price. A bidirectional head over N tokens that has no circulation, that is a reversible one, and that puts weight at least w_* on each token's successor needs a range of scores, the largest minus the smallest, of at least $\frac{1}{2}[(N-2)\log W_* - \log N]$,

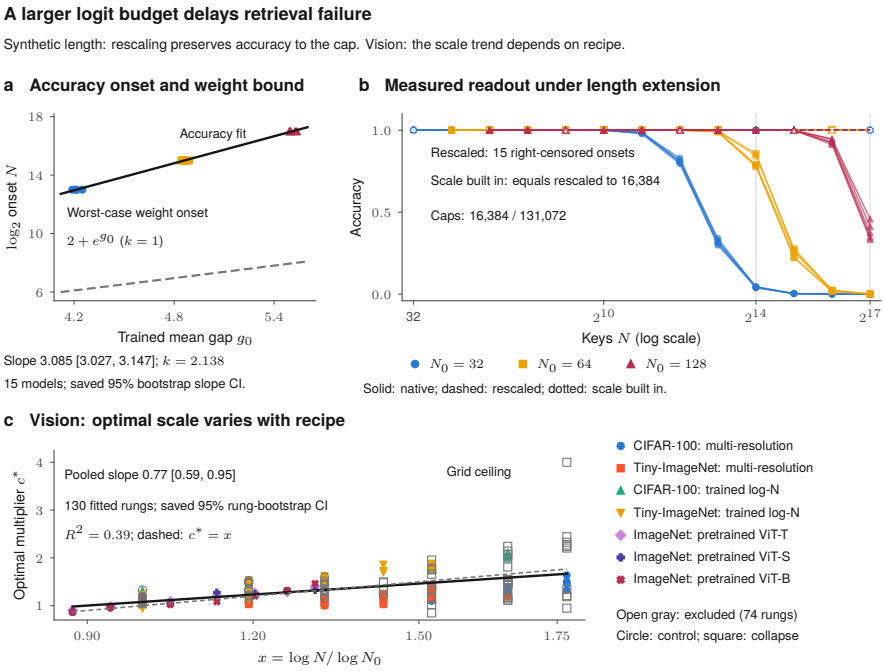


Figure 9.9: **Retrieval accuracy fails later at larger trained gaps, and log-length rescaling preserves it to the synthetic evaluation cap; vision's optimal scale has a recipe-dependent trend.** **(a)** The 15 native accuracy onsets of the length test (five seeds at each training length $N_0 = 32, 64, 128$), defined by the first doubling-grid value with cosine nearest-content decoding accuracy below 0.5. The trained gap g_0 is the query-mean partner logit minus the largest competing logit, excluding self. The black line uses the saved fit $\log_2 \widehat{N} = 3.0846 g_0 + 0.0028$; $k = 2.1381$ is its exponential rate. The dashed $k = 1$ comparator is the *worst-case weight onset* $2 + e^{g_0}$, not an observed weight threshold or an accuracy certificate for a mean gap. **(b)** Every seed's native (solid) and evaluation-rescaled (dashed) readout, with logits multiplied by $\log N / \log N_0$. All 15 rescaled onsets are right censored at 131,072; survival beyond that cap is unobserved. The first evaluation, to 16,384 keys, is an exact prefix of the evaluation to 131,072, so the two marked caps describe one set of 15 models. Dotted curves show an arm with the scale built into the forward pass, tested only to 16,384; its factor is one at the training length, so it has the same 15 models' weights and repeats the rescaled readout. All 45 seed lines are retained; near-coincident curves overlap, and rescaled markers are spaced along the recorded vertices. **(c)** Saved, sub-grid-refined optimal logit multipliers: 30 from-scratch models (two datasets, three recipes, five seeds), whose higher resolutions are upsampled test images, and three pretrained ViTs, whose inputs used ImageNet normalization where their checkpoints expect a mean and standard deviation of 0.5. The pooled fit uses 130 graceful rungs from graceful sources; dashed $c^* = x$ is a reference. Open gray squares show 44 excluded collapse rungs; open circles show 30 non-collapsed rungs excluded because their source uses single-resolution training. One excluded optimum reaches the multiplier grid ceiling $c = 4$. Graceful means native accuracy at least 0.4 times in-distribution accuracy and greater than twice chance. Displayed intervals are the saved descriptive 95% bootstrap slope intervals: 5,000 model resamples for (a) and 10,000 rung resamples for (c), not a model-level interval for (c) or a pointwise confidence band. The registered probe of the low-optimum collapse regime remains inconclusive.

which grows linearly in N when $w_* > \frac{1}{2}$ (a corollary of Theorem 10.1(c), not yet checked in Lean). The evaluation rescale above supplies the logarithmic growth, since it multiplies every logit by $\lambda = \log N / \log N_0$. That is the tempered rule of Section 9.2 in a model that has no null. Once a null is present, tempering multiplies it too, and with separated margins it then keeps both halves (Theorem 9.6); in the pilot it kept both empirically over the tested lengths, without the certificate (Section 17.2). The results of this section are the retrieval half of the law; Section 17.2 and Section 8.5 give its abstention half.

9.7 What holds in trained networks

Exact under premises checked network by network. The exact simultaneous budget (Section 9.1, with its retrieval half measured in Section 9.6): exact for any softmax row whose logits lie in $[-\tau, \tau]$: the bound is worst case, and τ is read off the network. The two-sided law and its tempered rule (Sections 9.1 and 9.2): exact for any softmax row with content logits in $[-\tau, \tau]$ and a null shared by absent and present queries, at any layer; tempering needs strictly separated margins, checked on the head. The conservation of routing information (Section 9.5): each part is exact within its class: Fano's bound and the order bound for any decoder; the expression bound for any soft majority; the memory floor for one-pass states with arbitrary later queries over unstructured items. The key-channel law assumes random keys on a sphere.

Shown in an idealized setting. Sums need a $\log N$ schedule at a fixed logit scale while maxima need none against backgrounds bounded below their threshold, and no router beats the argmax on the same scores (Sections 9.3 and 9.4; the exact condition is part (b) of Proposition 9.10): the argmax bound needs a uniformly placed target and a joint likelihood that depends on the target only through a nondecreasing factor of its own score; independent backgrounds with a monotone likelihood ratio satisfy this, exchangeable ones need not. The onset law assumes independent keys; the tests are constructed.

Chapter 10

An arrow: circulation and order

Chapter 3 split every head with a stationary law into a preference, a reciprocal exchange and a circulation, and every attention asymmetry into a sink and a cycle (Theorems 3.1 and 3.2); it also showed that a statistic unchanged by time reversal measures how much circulation there is, never its sign (Section 3.5). This chapter asks what the circulation is for. The answer is the sixth of the conditions of Chapter 0 (Section 0.3.2), the arrow: a head that cannot circulate cannot route reliably around a cycle, and can point along a sequence only by piling its stationary mass toward one end. Section 10.1 states the limits for a single head, and Section 10.2 the interventions that change circulation while fixing the stationary law and the reciprocal traffic. Sections 10.3 to 10.5 measure how much circulation trained encoders carry, how much their predictions depend on its orientation, and what it computes; Section 10.6 holds the place of a test of direction by composition. Section 10.7 prices order for routers that read a store written over time, and Section 10.8 says what of all this applies to decoders, which take their arrow from the mask.

10.1 Direction requires circulation

Circulation is how a single head routes. The following limits hold for row-stochastic kernels under the stated assumptions, softmax or not; for softmax attention, r is the cycle term of the scores (Theorem 3.2).

Theorem 10.1 (Direction requires circulation)**.** *Let P have a fixed positive stationary law π. Let $i_1, \dots, i_N$ ($N \geq 3$, indices cyclic) be distinct tokens, and write $p_t = P_{i_t i_{t+1}}$ (successor) and $q_t = P_{i_t i_{t-1}}$ (predecessor). In part (a), assume $p_t, q_t > 0$ for every t.*

(a) Cycle bound. $\sum_t \operatorname{logit}(p_t) \leq \mathcal{A} := \sum_t \log(p_t/q_t)$, *with equality when every row puts all its mass on its two neighbors; for softmax attention $\mathcal{A} = \sum_t r_{i_t i_{t+1}}$. If P is circulation-free, then $\mathcal{A} = 0$ (Kolmogorov's criterion; Kolmogoroff, 1936; Kelly, 1979). The successor-to-predecessor odds then have geometric mean* exactly *one, and the successor probabilities have geometric mean at most $\frac{1}{2}$. Routing with $p_t \geq 1 - \varepsilon$ on every row requires $\mathcal{A}/N \geq \log \frac{1-\varepsilon}{\varepsilon}$.*

(b) Reversible heads swap but cannot shift. *A circulation-free permutation kernel is an involution: reversible attention can exchange pairs of tokens but cannot rotate a cycle. Reversibility is not closed under composition. Two reflections, each symmetric, compose to a shift, so two circulation-free layers can do what one cannot.*

(c) Sinks are the price of reversible routing. *On an open sequence $1, \dots, N$, if P is circulation-free and $P_{i,i+1} \geq 1 - \varepsilon$ for $i < N$, then $\pi_1/\pi_{N-1} \leq \big(\varepsilon/(1-\varepsilon)\big)^{N-2}$. Such heads exist, for example*

by taking the softmax of $s_{ij} = -\beta(|i-j|-1)^2 + c\,j$ with $\beta > c > 0$ and scaling β and c together sufficiently for the chosen ε (symmetric scores plus a key bias).

(d) Circulation and positional chirality are different things. *A network whose scores depend only on token content and on $|i-j|$, with no absolute positions and with all other operations shared tokenwise, is equivariant to reversing token order whether or not its heads circulate: it can follow content-defined cycles but cannot tell left from right. Conversely, tied rotary attention has symmetric scores and $J = 0$ on every input, yet is not reflection-equivariant. It is chiral without circulating.*

(e) Gram scores prefer the reading token. *If $s_{ij} = \lambda \phi_i^\top \phi_j + b_j$ with $\lambda \geq 0$, for features ϕ and a key bias b (tied $Q = K$, with ϕ_i the query of token i; tied rotary attention, with the rotated queries), then around every cycle*

$$\sum_t \log \frac{p_t}{P_{i_t i_t}} = \sum_t \log \frac{q_t}{P_{i_t i_t}} = -\frac{\lambda}{2} \sum_t \|\phi_{i_t} - \phi_{i_{t+1}}\|^2 \leq 0. \tag{10.1}$$

On geometric average around any cycle, such a head attends to each neighbor at most as much as to the reading token itself. Symmetric scores with an indefinite form, such as $\frac{1}{2}(QK^\top + KQ^\top)$, carry no such restriction.

The proof is in Section A.11. Part (a) turns the familiar intuition that "symmetric attention cannot route" into an equality. It is stronger than the geometric-mean bound, and it gives the score budget required for accurate routing. For bilinear scores that budget is $2\sum_t x_{i_t}^\top W_{\mathrm{sk}} x_{i_{t+1}}$. Part (c) explains why a positional head in a text encoder need not circulate to point forward. On a line it can instead build a sink, although at an exponential cost in stationary mass. Whether trained heads choose the sink or the cycle is an empirical question, which Section 10.3 answers. Part (b) and the per-layer nature of all five statements matter for interpreting networks. The limits constrain what a single head can route, not what a network can compute from what its heads read (Section 10.5). Part (d) is a caution in the other direction: a head with $J = 0$ can still be sensitive to order through other means, and circulation measures one specific, nonreversible form of directional computation. Part (e) separates circulation-free designs from one another. Every circulation-free head reads both neighbors of a cycle equally in the geometric mean. Gram scores cannot read them more than they read the token itself; an indefinite form can.

The statement of Theorem 10.1 is per head. Reversibility is not closed under composition: two reflections compose to a shift, and a network can compute direction from reads of both neighbors. On the synthetic cyclic-retrieval task (small models, means over five seeds), one symmetrized layer with indefinite scores followed by an MLP reaches 96.1% successor retrieval (98.2–98.5% at depth 2–4), while tying queries to keys lags (22%, 60% and 76% at depth 1, 2 and 4), because, on geometric average around any cycle, Gram scores cannot prefer a neighbor to the token itself (Theorem 10.1(e); the experiment is Section 10.5). Decoders take their arrow from the mask (Section 10.8). Under it, tied (Gram) scores still copy less surely than untied scores when the offset varies in a small synthetic two-layer test (five seeds; mean per-seed cross-entropy ratio 3.4; accuracy 0.8 points lower), and without a mask they are at parity on bidirectional lookup: a cost of the Gram form, not a removal of the arrow, measured on trained models (Bruce, 2026d).

10.2 Canonical interventions on circulation

The decomposition would be a descriptive curiosity if circulation could not be manipulated in isolation. It can, and the natural manipulations turn out to be canonical in a precise sense. The first is the matched path P_α, which is the sweep (1.3) of Section 1.2 with the dose α in the place of t.

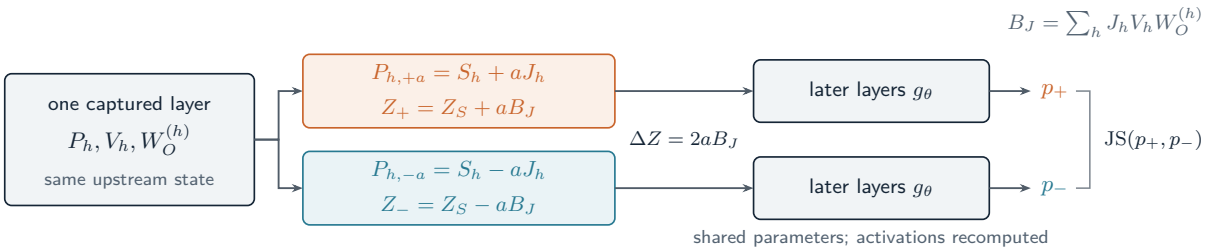


Figure 10.1: **The matched reversal experiment.** All heads of one captured layer are replayed from the same upstream state with $P_{h,\pm a} = S_h \pm aJ_h$ (Proposition 10.2). The two arms share every invariant listed in the figure, and their attention outputs differ by exactly $2aB_J$ (10.2). Later layers recompute, so the predictions $p_\pm$ need not respond linearly. The paired divergence $\mathrm{JS}(p_+, p_-)$ measures how much the model uses the orientation of its circulation.

Proposition 10.2 (Matched circulation path)**.** *For real α let $P_\alpha = S + \alpha J = \frac{1+\alpha}{2}P + \frac{1-\alpha}{2}P^\dagger$. For $|\alpha| \le 1$, P_α is a nonnegative stochastic kernel with stationary law π. Nonnegativity in fact holds exactly for $|\alpha| \le \min_{i\neq j} \coth(\frac{1}{2}|\log(F_{ij}/F_{ji})|)$, a bound that is always at least one. Along the whole path the stationary law π, the reciprocal traffic A, every stationary cut flow $\sum_{i\in U, j\notin U} \pi_i (P_\alpha)_{ij}$ and the Dirichlet form are fixed, while the current is αC. Opposite doses are adjoint, $P_{-a} = P_a^\dagger$, and therefore share their eigenvalues, their singular values in stationary coordinates, their current norm, their entropy production and their stationary conditional entropy.*

The proof is in Section A.11. The symmetry between opposite doses extends to every amplifier of the operator: for any normalized amplifier f of P_α, aligned readouts are even in α, while cross readouts and eigenvectors respond at first order (Theorem 3.4). We call $\alpha = 0$ *stilling* and $\alpha = -1$ *reversal*. The pair $P_{\pm a}$ (Figure 10.1) is the controlled contrast used throughout. Its two members differ only in the orientation of circulation, so any difference in a model's behavior between them is attributable to that orientation. Stilling changes the magnitude of circulation as well, and is therefore a different comparison. Nothing in the pair is matched by construction *downstream* of the operator; that is what the experiments measure.

Many practical operations remove circulation: tying queries to keys, symmetrizing scores, replacing a head by a symmetric kernel. Theorem 10.3 shows that among all of them, stilling and score symmetrization are distinguished. They are the two information projections onto circulation-free attention (Csiszár, 1975; Amari & Nagaoka, 2000). Wolfer & Watanabe (2021) study both projections abstractly for reversible chains; the statements below identify them for attention and give the softmax form. For kernels R, Q with stationary law μ of R, write $D(R\|Q) = \sum_i \mu_i D_{\mathrm{KL}}(R_{i\cdot}\|Q_{i\cdot})$ for the *stationary divergence rate*.

Theorem 10.3 (Stilling and symmetrization are canonical)**.** *(a)* Stilling is the m-projection. *For every circulation-free kernel R, with any stationary law, that is positive wherever A is,*

$$D(P\|R) = D(P\|S) + D(S\|R),$$

where the divergences are weighted by π. Hence S is the unique minimizer of $D(P\|R)$ over all circulation-free kernels. The information cost of any *circulation-free replacement of a head splits exactly into circulation removed, $\mathcal{I}(P) = D(P\|S)$, plus reciprocal structure changed,*

$D(S\|R)$. Along the path, $D(P_\alpha\|S) = H_\pi(S) - H_\pi(P_\alpha)$ is even and convex in α, so stilling maximizes stationary conditional entropy.

(b) Symmetrization is the e-projection. *Assume $G_{ij} = \sqrt{P_{ij}P_{ji}}$ is irreducible, with Perron root $\rho \le 1$ and positive Perron vector v. Then $E_{ij} = G_{ij}v_j/(\rho v_i)$ is circulation-free with stationary law proportional to v^2, and for every circulation-free R,*

$$D(R\|P) = D(R\|E) + D(E\|P), \qquad D(E\|P) = -\log\rho,$$

with $\rho = 1$ exactly when P is itself circulation-free. For softmax attention, $E = \mathrm{softmax}(\frac{1}{2}(s + s^\top) + \mathbf{1}\beta^\top)$ with $\beta_j = \log v_j - \frac{1}{2}\log z_j$ and $z_j = \sum_k e^{s_{jk}}$. Replacing the scores by their symmetric part and refitting a key bias is exactly the reverse information projection. Unlike stilling, it generally moves the stationary law.

The proof is in Section A.11. The theorem converts an informal question, whether an architectural change removes "the directional part" of a head, into an exact accounting. Weight tying and score symmetrization (Courtois et al., 2024; Kowsher et al., 2025; Kayyam et al., 2026) remove the circulation of every head by construction, and part (b) identifies the cheapest such kernel in the reverse divergence. Part (a) says the edit that removes circulation and nothing else, in the forward divergence, is unique and is the stilling operation.

A practical obstacle to intervening on circulation is that fused attention kernels never materialize P (Dao et al., 2022). Reversal does not need it to.

Proposition 10.4 (Stationary reversal as attention)**.** *Let $P = \mathrm{softmax}(\lambda QK^\top)$ and write $\ell_i = \log\sum_j e^{\lambda q_i\cdot k_j}$ for the row log-normalizers returned by fused attention. Then*

$$P^\dagger = \mathrm{softmax}(\lambda KQ^\top + \mathbf{1}b^\top), \qquad b_i = \log\pi_i - \ell_i.$$

The swapped call, with keys as queries, queries as keys and key bias b, returns log-normalizers $\ell'_j = \log(\pi^\top P)_j$, so $\ell'_j - \log\pi_j = \log(P^\dagger\mathbf{1})_j$. The reversed row sums are its own stationarity certificate. Iterating $\log\pi \leftarrow \log\mathrm{softmax}(\ell')$, which uses only the normalizers, is power iteration for π. Hence $P_\alpha V = \frac{1+\alpha}{2}\,\mathrm{Attn}(Q,K,V) + \frac{1-\alpha}{2}\,\mathrm{Attn}(K,Q,V;b)$ costs two fused attention calls and $O(N)$ memory.

The proof is in Section A.11. We verified the construction with FlexAttention (Dong et al., 2025) against the dense computation, with errors of 10^{-14} in `float64` (Bruce, 2026g, Appendix C). One warning from building this check: plain power iteration stalls on nearly periodic heads, such as the near-permutation heads of Section 10.5. The lazy chain $\frac{1}{2}(I + P)$ has the same π and avoids this, and the certificate should always be checked. For dense analysis we use the subtraction-free elimination of Grassmann et al. (1985) (GTH). It retains relative accuracy for stationary masses as small as 10^{-175}, which do occur (Section 10.3).

An operator edit matters only through the content it delivers. For heads h with values V_h and output-projection blocks $W_O^{(h)}$, the attention output under a matched edit of all heads at one layer is exactly affine in the dose:

$$Z(\alpha) = Z_S + \alpha B_J, \qquad B_J = \sum_h J_h V_h W_O^{(h)}, \tag{10.2}$$

before any downstream recomputation. B_J is invariant to reparametrizations $V_h \mapsto V_h T_h$, $W_O^{(h)} \mapsto T_h^{-1}W_O^{(h)}$. It can be small even when J is large, because values may annihilate the current and heads may cancel. The downstream network can also be insensitive to it. A large current, a large delivered update, a large change in predictions and a *beneficial* change are four different findings, and the analysis keeps them apart.

Proposition 10.5 (Signed and label-free response)**.** *Let $\ell(\alpha)$ be a scalar outcome (a loss or logit margin) after downstream recomputation from $Z(\alpha)$, smooth near Z_S.*

(a) Odd response. $o(a) = \frac{1}{2}[\ell(a) - \ell(-a)] = \frac{1}{2}\int_{-a}^{a}\langle \nabla_Z \ell(Z_S + tB_J), B_J\rangle\, dt$ *exactly. The unfitted midpoint prediction $\hat{o}(a) = a\langle \nabla_Z \ell(Z_S), B_J\rangle$ has error at most $M_3 a^3/6$. With the two endpoint gradients, Simpson's rule has error at most $M_5 a^5/180$. Here M_k bounds the k-th derivative along the path.*

(b) Label-free divergence. *For predictive distributions $p_{\pm a}$ with logits z,* $\mathrm{JS}(p_{+a}, p_{-a}) = \frac{a^2}{2}\, g^\top H_0 g + O(a^4)$*, where $g = Dz(Z_S)[B_J]$ is one forward-mode Jacobian–vector product and the categorical Fisher metric is $H_0 = \mathrm{diag}(p_0) - p_0 p_0^\top$ (Martens, 2020). Writing $g^\top H_0 g = \Phi_{\mathrm{Fisher}} \|B_J\|_F^2$ factors the response into* delivered energy $\|B_J\|_F^2$ *and* Fisher gain Φ_{Fisher}*, the sensitivity of the prediction per unit of delivered energy.*

(c) Readout visibility. $\pi^\top P_\alpha = \pi^\top$*, so a linear readout that pools a head's output with weights π cannot see the edit. A class-token readout sees only the class-token row,* $(P^\dagger)_{\mathrm{cls},j} = \pi_j P_{j,\mathrm{cls}}/\pi_{\mathrm{cls}}$*: reversal exchanges what the class token reads with what reads it.*

The proof is in Section A.11. Part (a) is attribution patching along a specified, matched direction (Kramár et al., 2024). The odd contrast cancels all even-order terms, and the exact integral is a path attribution (Sundararajan et al., 2017; Hanna et al., 2024). None of this is a new attribution principle. What is new is the pairing with an intervention whose invariants are known exactly, which turns an explanation of circulation into a falsifiable, signed prediction. Part (b) is what makes the specificity result of Section 10.4 interpretable. Two edits of equal operator magnitude can differ in predictive effect only by delivering different energy or by aiming it differently relative to the Fisher geometry. *Validity.* Stationary semantics require certification on the captured matrix. For any positive reference q, $(D_q^{-1} P^\top D_q \mathbf{1} - \mathbf{1})_i = (q^\top P - q^\top)_i / q_i$, so a small absolute stationarity residual can produce a large reversal error at a token of tiny mass. We therefore require componentwise stationarity and reversed row sums to within 10^{-9}, together with nonnegativity and exact identity replay. Clamping a stationary mass or adding teleportation changes the operator, and we never do either. Every kernel analyzed with these semantics in Sections 10.3 to 10.5, 184,320 in the census and all intervention sites, passes these checks.

10.3 A census of circulation

How much circulation do trained transformers carry, and where? We decomposed every head at every layer of five public encoders. Three are vision transformers, trained with different recipes and position encodings: DeiT-B (Touvron et al., 2021), a rotary-position ViT-B (Heo et al., 2024) and ViT-B (Dosovitskiy et al., 2021). Two are language encoders: BERT-base (Devlin et al., 2019) and RoBERTa-base (Liu et al., 2019). Inputs were 256 randomly drawn ImageNet validation images (Deng et al., 2009) or WikiText-2 validation lines of 16–126 RoBERTa tokens (Merity et al., 2017). Models ran in `float64`, stationary laws were computed by GTH, and every kernel was certified. All 184,320 head kernels pass, with componentwise stationarity and reversed-row residuals below 6×10^{-15}. The pipeline reproduces an earlier independent measurement: median head-mean ratios $\|\widetilde{J}\|_F / \|\widetilde{S}\|_F$ at layer 6 are 0.518, 0.541 and 0.387 for DeiT, RoPE-ViT and ViT, against 0.516, 0.540 and 0.383 on a disjoint image bank. Figure 10.2 and Table 10.1 summarize the results; the protocol is in Bruce (2026g, Appendix D), and additional views close this section.

Circulation is a mid-network phenomenon in vision. In all three vision transformers the first two layers are nearly circulation-free, with median κ_{circ} between 0.01 and 0.04. Circulation rises through

Table 10.1: **Circulation census.** Every head at every layer of five pretrained encoders, 256 inputs per model (ImageNet validation images or WikiText-2 validation lines). All kernels pass the stationary certificate (componentwise stationarity residuals and reversed-row-sum errors $\leq 10^{-14}$). κ_{circ}: median circulation index at the first layer, at its peak layer and at the last layer. Heads $\kappa_{\mathrm{circ}} > \frac{1}{2}$: fraction of the 144 heads whose median index exceeds one half. Cycle share: fraction of the score asymmetry $\|a\|_F^2$ carried by the cycle term (Theorem 3.2), at layer 0 and at its peak. Hub: median enrichment of current on edges incident to special tokens relative to their share of edges (layers 1–10). Min. mass: smallest certified stationary mass.

Model	Kernels	κ_{circ} L0	κ_{circ} peak	κ_{circ} last	Heads $\kappa_{\mathrm{circ}} > \frac{1}{2}$	Cycle share	Hub	Min. mass
DeiT-B	36,864	0.02	0.38 (L8)	0.08	13%	5 / 51%	22×	4×10^{-10}
RoPE-ViT-B	36,864	0.03	0.42 (L8)	0.20	23%	8 / 75%	24×	3×10^{-175}
ViT-B	36,864	0.01	0.30 (L7)	0.16	11%	4 / 43%	23×	4×10^{-27}
BERT-base	36,864	0.26	0.68 (L5)	0.12	36%	59 / 59%	9×	4×10^{-7}
RoBERTa-base	36,864	0.22	0.58 (L8)	0.20	41%	70 / 70%	10×	2×10^{-58}

the middle of the network to a peak of 0.30–0.42 at layers 7–8 and falls again toward the classifier. Individual heads reach $\kappa_{\mathrm{circ}} > 0.99$, close to the ceiling that Theorem 3.1 allows. The language encoders circulate more. Median κ_{circ} peaks at 0.68 in BERT (layer 5) and 0.58 in RoBERTa (layer 8), and 36–41% of their heads have a median index above one half, against 11–23% in the vision models. The rotary-position model carries the most circulation among the vision transformers, including at its final layer (0.20 versus 0.08 and 0.16).

Asymmetry is sink early in the network and cycle in the middle. The score-level split of Theorem 3.2 shows what the asymmetry is made of. At layer 0 of the vision transformers the cycle term carries only 4–8% of the score asymmetry; the rest is a key-bias potential, a pure preference that cannot circulate. By the middle layers the cycle term carries 43–75%. In the language encoders cycles already dominate at the first layer (59–70%). The cycle's share of total row-centred score energy tracks the operator-level index closely, with Spearman correlations of 0.82–0.90 across vision kernels and 0.46–0.57 across text kernels. It can therefore serve as a solver-free screen.

Positional heads circulate; they do not build sinks. BERT has eight heads that put more than half of their attention on the previous (four heads) or the next (four heads) token. They lie in the layers identified by Clark et al. (2019): layers 2, 4, 7 and 8 for previous-token heads in their one-based indexing. These heads are near-maximal circulators, with median $\kappa_{\mathrm{circ}} = 0.95$ against 0.32 for the other heads. The three such heads of RoBERTa also have median $\kappa_{\mathrm{circ}} = 0.95$, against 0.40. Theorem 10.1(c) gives the circulation-free alternative when the successor probability has a common lower bound $1 - \varepsilon > \frac{1}{2}$ at every nonterminal token: an exponential boundary sink, with $\log \pi$ rising by at least $\log \frac{1-\varepsilon}{\varepsilon}$ per interior step. For a uniform successor bound near 0.9, that is more than two nats per interior step. The trained heads choose the cycle instead. Their fitted log-slopes along content positions have per-head medians between -0.014 and $+0.055$ per token, and their current returns through the special tokens and through low-weight leakage (Table 10.2). The canonical "directional" heads of interpretability are, quantitatively, circulation. *Sinks are circulation hubs.* Preference and circulation are orthogonal components, but they are coupled in trained models. The tokens at which attention accumulates are also where cycles close. Edges incident to the class token carry 22–24 times their share of the stationary current in the vision transformers, and edges incident to [CLS]/[SEP] carry 9–10 times their share in the language encoders (Figure 10.2d). This fits Theorem 10.1(c). On an open sequence, a head whose successor probability has a common lower bound above one half at every nonterminal token must either pile up stationary mass exponentially

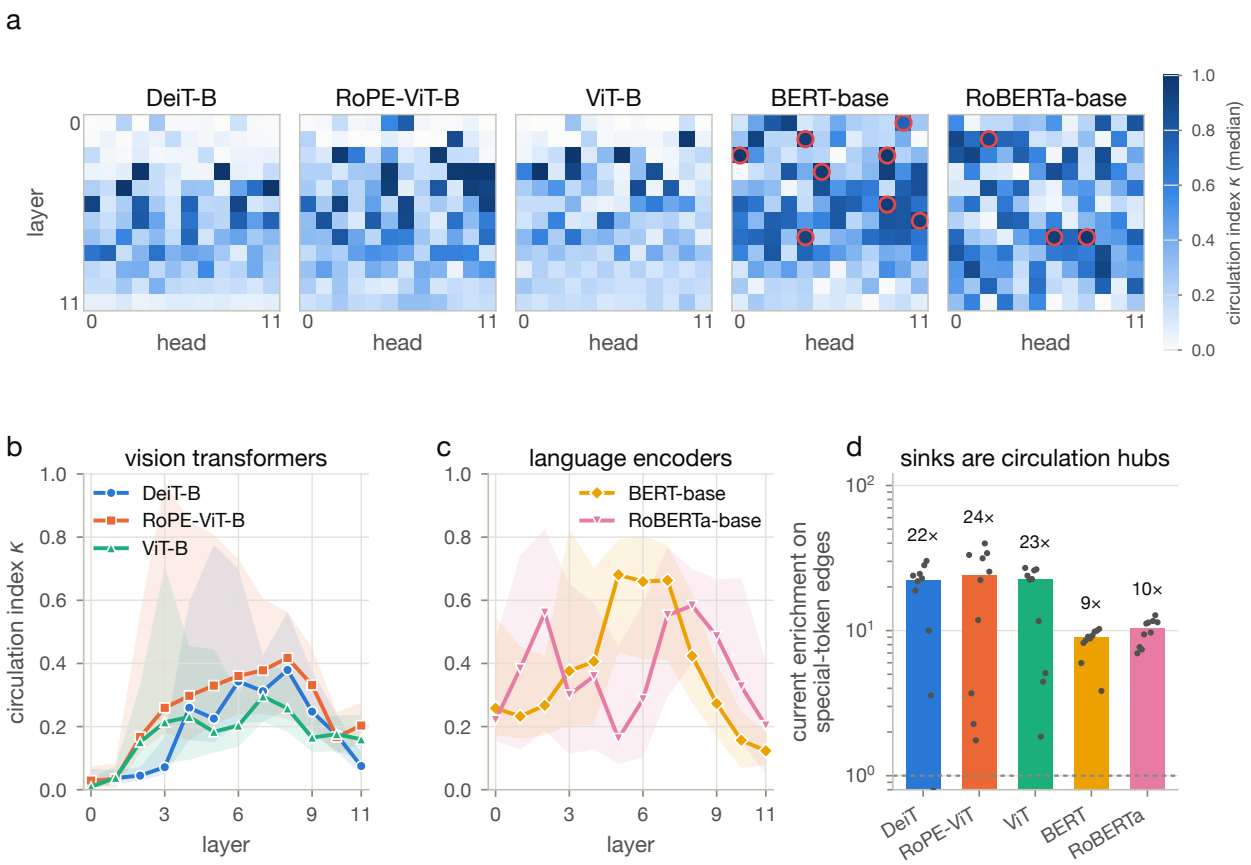


Figure 10.2: **Circulation in five pretrained encoders. a** Median circulation index κ_{circ} of every head (columns) at every layer (rows) over 256 inputs. Red circles mark the positional heads of the language encoders, those that put more than half of their attention on the previous or the next token. **b,c** Layer profiles: median and interquartile range over inputs and heads. **d** Enrichment of stationary current on edges incident to special tokens ([CLS] in vision; [CLS]/[SEP] or <s>/</s> in text), relative to the share of such edges. Bars show medians over layers 1–10 and dots show individual layers. The dashed line marks no enrichment. The panels write κ for the circulation index κ_{circ}.

or circulate, and a circulating current, being divergence-free, needs a return path. Sink tokens provide one.

Where the current lives in vision transformers. Figure 10.3 shows how each layer spends its operator energy and how much of its score asymmetry is cycle. The current shares and flow coherences below are medians over inputs and heads at each layer. Most stationary current runs along long-range patch–patch edges (45–96% by layer, rising with depth). Local edges between neighboring patches carry 3–32% of the current while forming 3.6% of all edges, an over-representation of up to ninefold in the early and middle layers. In DeiT and ViT, edges incident to the class token carry 12–52% of the current in layers 0–7, against 1% of edges, and little in the last layers, where the class token's stationary mass collapses (Table 10.3). RoPE-ViT routes little current through its class token in layers 2–4 (2–4%) and much in layers 7–11 (21–40%), matching its persistent class-token mass. The patch-level current field is directionally coherent in early and middle layers. The ratio of the norm of the summed flow vector to the summed norms of the per-patch flow vectors is 0.16–0.45

Table 10.2: **Positional heads are near-maximal circulators.** Heads that put more than half of their attention on the previous or next token (median over 256 lines; zero-based layer and head indices). κ_{circ}: circulation index. $\mathcal{I}/\log 2$: circulation information as a fraction of its maximum. Slope: least-squares slope of $\log \pi_i$ against position over content tokens, per token. For a circulation-free head whose successor probability has the common lower bound $1-\varepsilon > \frac{1}{2}$ at every nonterminal token, $\log \pi$ rises by at least $\log \frac{1-\varepsilon}{\varepsilon}$ per interior step (Theorem 10.1(c)), about two nats for $\varepsilon \approx 0.1$. The Frac. column is a median of per-line mean attention weights, not this pointwise lower bound. C special/adjacent: share of $\sum |C_{ij}|$ on edges incident to special tokens and between adjacent tokens (baseline edge shares are about 5% and 2.5%).

Model	Layer	Head	Type	Frac.	κ_{circ}	$\mathcal{I}/\log 2$	Slope	C special	C adjacent
BERT-base	0	10	next	0.54	0.752	0.42	+0.005	0.18	0.58
BERT-base	1	4	prev	0.57	0.951	0.53	-0.008	0.32	0.60
BERT-base	2	0	next	1.00	0.987	0.19	+0.028	0.09	0.91
BERT-base	2	9	next	1.00	0.990	0.83	+0.027	0.06	0.94
BERT-base	3	5	prev	0.76	0.974	0.41	+0.008	0.19	0.75
BERT-base	5	9	next	0.67	0.956	0.46	+0.002	0.25	0.62
BERT-base	6	11	prev	0.65	0.943	0.36	+0.002	0.23	0.68
BERT-base	7	4	prev	0.56	0.928	0.13	+0.002	0.27	0.56
RoBERTa-base	1	2	next	0.89	0.903	0.01	+0.055	0.05	0.88
RoBERTa-base	7	6	next	0.62	0.954	0.21	+0.008	0.33	0.63
RoBERTa-base	7	8	prev	0.83	0.989	0.91	-0.014	0.32	0.66

Table 10.3: Stationary class-token mass in the three vision transformers, summarized over the 512 images of the confirmatory study of Section 10.4 and the twelve heads of each cell. Heads are nested measurements, not independent image replicates. Uniform class-token mass is $1/197 \approx 0.00508$. Effective token count is $1/\sum_i \pi_i^2$.

Checkpoint	Layer	Median π_{CLS}	Mean π_{CLS}	Max. π_{CLS}	Median eff. tokens
DeiT	6	0.13832	0.19849	0.86600	32.60
DeiT	11	0.00488	0.00612	0.08712	44.01
RoPE	6	0.24412	0.24401	0.79123	14.76
RoPE	11	0.14786	0.17024	0.57525	29.18
ViT	6	0.15470	0.18907	0.71724	20.40
ViT	11	0.00314	0.00357	0.02386	30.29

in layers 0–7 of DeiT and ViT, and it falls to near zero in their final layers. RoPE-ViT differs: its coherence rises again in layers 6–9.

10.4 Learned circulation is consequential

The census shows that trained heads circulate. Does the model *use* the orientation of that circulation? And if so, is the learned orientation special, or would any circulation of the same size matter as much? The matched pair of Proposition 10.2, applied to all twelve heads of a layer at once in the three vision transformers of the census, answers both. At layer 6 of 12 the model is replayed from the same upstream state with all heads set to P_{+a} or P_{-a}, and the outcome is the paired Jensen–Shannon divergence between the two predictive distributions at half dose, $a = \frac{1}{2}$, which is label-free and sign-agnostic. The learned current is compared with feasible alternatives: currents

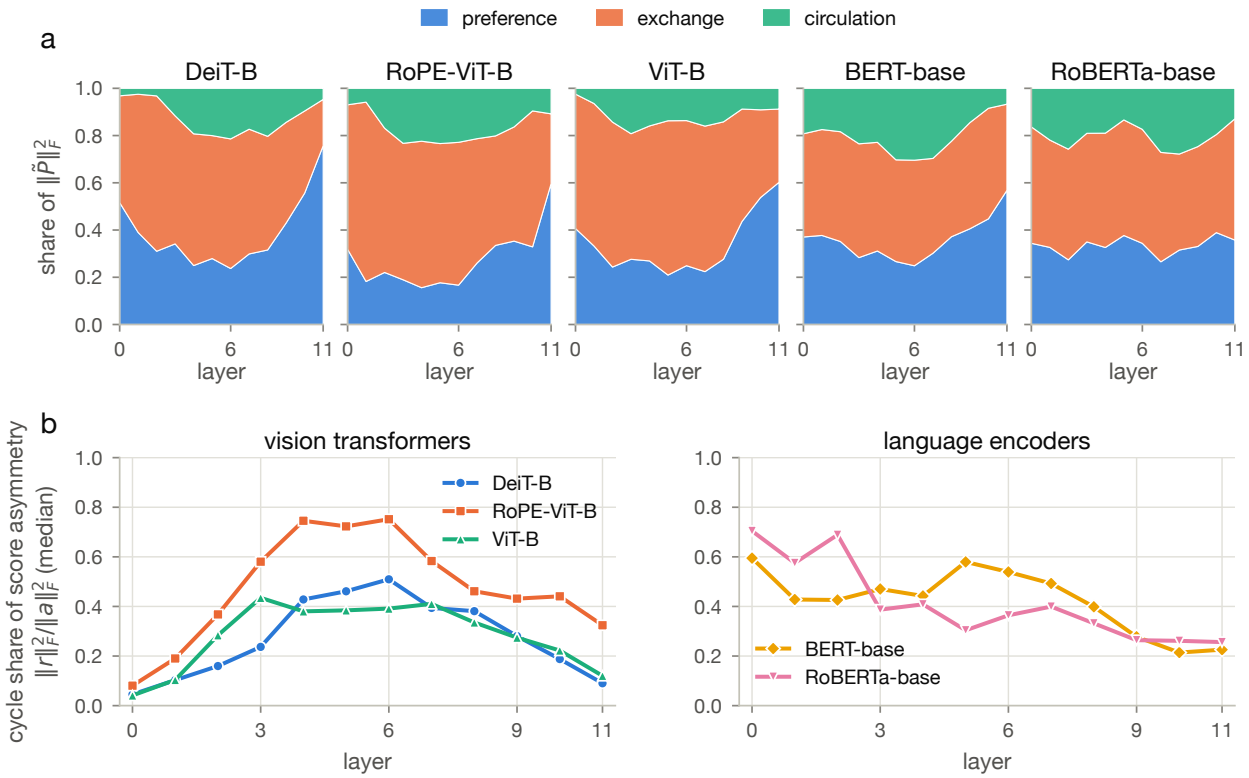


Figure 10.3: **Census, additional views. a** Mean shares of $\|\widetilde{P}\|_F^2$ carried by preference, exchange and circulation at each layer. Preference is U-shaped over depth, exchange dominates the middle of the network, and circulation peaks mid-network in vision and is broadly spread in text. **b** Share of the score asymmetry carried by the cycle term (the remainder is the key-bias sink potential). Early vision layers are almost pure sink.

C' that are skew, divergence-free and within capacity, $|C'_{ij}| \le 2A_{ij}$, so that each defines a valid kernel $D_\pi^{-1}(A \pm C'/2)$ with the same stationary law and the same reciprocal traffic as the learned head. Each is scaled to the learned current's canonical norm in every head, and twenty independent draws are reversed in exactly the same way (Bruce, 2026g, Appendix E). In a confirmatory study on 512 held-out images per model, with its protocol fixed before evaluation, reversing the learned current changes the predictive distribution 11.15, 7.94 and 12.18 times more than reversing matched random currents (Table 10.4). All three primary tests reach the resolution floor of the bootstrap, Holm-adjusted $p \approx 1.5 \times 10^{-4}$. The separation is not a tail artifact. The learned response exceeds the random mean on 497–510 of 512 images per model, and exceeds all twenty draws on 383–467 (Figure 10.4). The mean is heavy-tailed: the 5% of images with the largest contrasts contribute 51–62% of it. Removing them still leaves ratios of 8.2, 4.7 and 8.5. Monte Carlo error from using twenty draws is 0.2–0.3% of the effect. The ordering persists at the final layer (ratios 5.9–15.7) and at quarter dose, where halving the dose divides the paired JS by 3.8–4.0, as the quadratic law of Proposition 10.5(b) predicts. Opposite orientations produce different predictions even though they share preference, reciprocal traffic, cut flows, Dirichlet form and spectrum. Those invariants therefore do not determine the model's response; the orientation of circulation does.

Why: learned circulation is aimed. Two currents of equal operator magnitude can differ in predictive effect in only two ways (Proposition 10.5(b)). They can deliver different amounts of content through the values and output projection, or they can aim it differently relative to the

Table 10.4: **Learned versus feasible random circulation** (confirmatory study; 512 held-out images per model). Paired JS at half dose, in 10^{-3} nats. Random JS is averaged over 20 matched draws within each image. Brackets give 95% image-bootstrap intervals. "Above" counts images whose learned JS exceeds the mean of their random draws, and "all" counts images where it exceeds every draw. Layer 6 is the primary site (Holm-adjusted one-sided $p = 3/20{,}001$ for each model, the resolution floor); layer 11 is secondary.

Model	Layer	Learned JS	Random JS	Ratio of means [95% CI]	Above	All 20
DeiT-B	6	5.484	0.492	11.15 [9.53, 13.07]	508	467
RoPE-ViT-B	6	2.874	0.362	7.94 [6.44, 9.68]	497	383
ViT-B	6	3.012	0.247	12.18 [9.87, 15.09]	510	462
DeiT-B	11	1.773	0.113	15.69 [12.07, 20.34]	506	–
RoPE-ViT-B	11	3.014	0.235	12.83 [11.09, 15.07]	511	–
ViT-B	11	0.999	0.168	5.95 [5.11, 6.95]	494	–

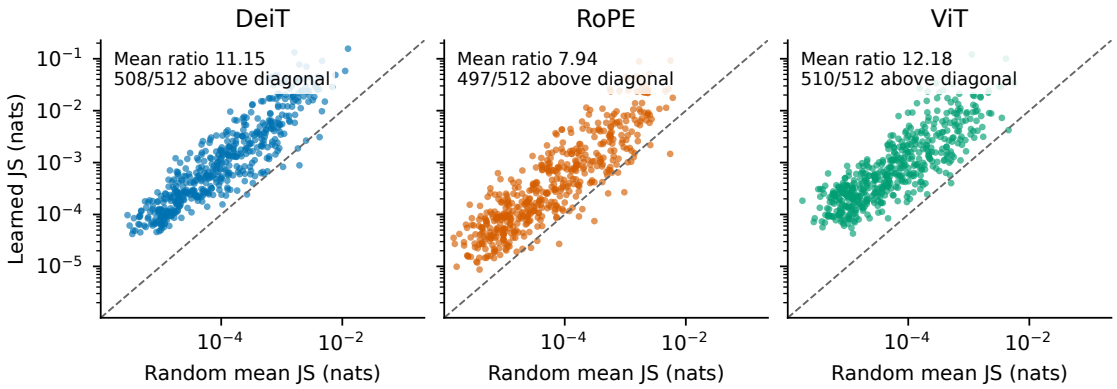


Figure 10.4: **Learned circulation is unusually consequential on nearly every image.** Each point is one of 512 held-out images of the confirmatory study at layer 6 and half dose. It compares the paired JS of reversing the learned current (vertical) with the mean over 20 feasible random currents of equal operator magnitude (horizontal). Both axes are logarithmic with identical limits. The diagonal marks equal response. Every image is shown.

network's sensitivity. We measured both. On a separate bank of 128 images per model, disjoint from the census bank, we reimplemented the feasible-current sampler (Bruce, 2026g, Appendix C). Per image, 42 operator directions were built at identical per-head canonical norm, equal to half the learned current's canonical norm $\|\widetilde{J}_h\|_F/2$: the learned current $J/2$; twenty feasible random currents, each head-wise draw reaching the target before being scaled down, and none ever scaled up; the learned exchange $S - \mathbf{1}\pi^\top$, rescaled per head; and twenty feasible random exchanges, whose flow perturbations are symmetric, zero-margin and within capacity, rescaled per head. Random currents and random exchanges come from the same projection scheme, so the two controls are constructed symmetrically. For every direction we computed the delivered update and the label-free Fisher prediction $\frac{1}{2}g^\top H_0 g$ from one forward-mode Jacobian–vector product; for circulation directions we also computed the replayed paired JS (Table 10.5). The fresh bank reproduces the confirmatory contrast. Learned/random ratios are 12.9, 8.8 and 13.8 for DeiT, RoPE-ViT and ViT, against 11.2, 7.9 and 12.2 in the confirmatory study. The learned response exceeds the random mean on

Table 10.5: **Mechanism of specificity** (layer 6, half dose, 128 fresh images per model, 20 draws per family). Ratios are learned/random ratios of means with 95% image-bootstrap intervals. *Observed*: replayed paired JS. *Fisher*: label-free prediction $\frac{1}{2}g^\top H_0 g$. *Fisher* = *energy* × *gain* exactly. *Exchange*: the same Fisher ratio for learned versus random exchange at identical canonical norm. *Circ./exch.*: circulation specificity divided by exchange specificity.

Model	Observed	Fisher	Energy	Gain	Exchange	Circ./exch.
DeiT-B	12.94 [9.01, 18.11]	12.51 [8.97, 17.12]	2.67 [2.60, 2.74]	4.69 [3.40, 6.38]	6.45 [3.75, 10.41]	1.94 [1.28, 3.27]
RoPE-ViT-B	8.78 [6.44, 11.37]	8.99 [6.54, 11.55]	2.31 [2.26, 2.35]	3.90 [2.87, 5.03]	5.11 [3.42, 7.44]	1.76 [1.01, 2.95]
ViT-B	13.76 [10.67, 18.08]	13.77 [10.85, 17.62]	3.79 [3.69, 3.90]	3.63 [2.88, 4.64]	8.37 [6.18, 11.14]	1.64 [1.22, 2.24]

Model	>mean	>all	ρ_F	MRE_L	MRE_R	gain>1	circ>exch	F_J/F_E	$\lvert\cos\rvert$ L / R (10^{-3})
DeiT-B	128/128	119/128	0.992	5.6%	2.6%	97%	77%	1.75	3.50 / 1.48
RoPE-ViT-B	127/128	104/128	0.998	5.5%	4.2%	95%	66%	1.21	2.58 / 1.37
ViT-B	128/128	123/128	0.998	11.2%	2.0%	98%	76%	1.41	4.25 / 1.64

>mean, >all: images on which learned JS exceeds the mean of, or all of, the 20 random draws. ρ_F: Spearman correlation between predicted and observed JS over all learned and random directions. MRE: median relative error of the Fisher prediction. gain>1: images on which the learned direction has higher Fisher gain than the geometric mean of the random draws. circ>exch: images on which circulation specificity exceeds exchange specificity. F_J/F_E: Fisher response of learned circulation over learned exchange at equal canonical norm. $\lvert\cos\rvert$: median absolute cosine between delivered update and loss gradient, learned (L) and random (R) circulation.

127–128 of 128 images, and exceeds all twenty draws on 104–123. The Fisher formula predicts the paired divergence of individual directions, learned and random alike, across five orders of magnitude (Spearman 0.992–0.998; Figure 10.5a). Median relative errors are 6–11% for learned and 2–4% for random directions, and the formula reproduces the ratio itself (predicted 12.5, 9.0 and 13.8). The factorization $g^\top H_0 g = \Phi_{\text{Fisher}}\|B\|_F^2$, with B the update a direction delivers, then explains the ratio (Figure 10.5b). Learned circulation delivers 2.3–3.8 times the energy of random currents of the same operator magnitude. Because J has zero row sums, $(JV)_i = \sum_j J_{ij}(V_j - V_i)$: the learned current moves content between tokens whose values actually differ. And each unit of delivered energy moves the prediction 3.6–4.7 times more: it is aimed at directions to which the rest of the network is sensitive. Neither factor alone suffices. Delivered energy accounts for 39–45% of the log ratio on a typical image, and the learned direction has the higher Fisher gain on 95–98% of images. Consistent with aim, the median absolute cosine between the delivered current and the loss gradient is 1.9–2.6 times that of random currents. Is this specific to circulation, or is any learned attention structure more consequential than random structure of the same size? The same Fisher analysis applied to exchange compares the learned exchange with the twenty feasible random reciprocal perturbations. The Fisher-predicted response of learned exchange is 5.1–8.4 times that of random exchange, so part of the specificity reflects learned attention structure in general. But circulation is more specific than exchange by this measure in all three models (Figure 10.5c), by factors of 1.94 [1.28, 3.27], 1.76 [1.01, 2.95] and 1.64 [1.22, 2.24], on 66–77% of images. At equal canonical norm, learned circulation also has 1.2–1.75 times the Fisher-predicted response of learned exchange while delivering the same energy (0.95–1.05 times). Learned circulation is thus aimed, and more sharply aimed than the model's own reciprocal structure.

The delivered direction predicts the signed response. Proposition 10.5(a) turns an explanation of circulation into a signed, falsifiable prediction. The unfitted midpoint rule $\hat{o}(a) = a\langle \nabla_Z \ell(Z_S), B_J\rangle$, with true-label cross-entropy ℓ and no fitted coefficient, explains the learned odd loss at layer 6 with $R^2 = 0.993$, 0.990 and 0.991. It gets the sign right on 97.7%, 98.8% and 94.9% of images, with Spearman correlations of 0.99, 0.995 and 0.98 (Figure 10.6). At the final layer, $R^2 \geq 0.9994$ and sign agreement is at least 99.8%. The same rule predicts random-direction responses at least as

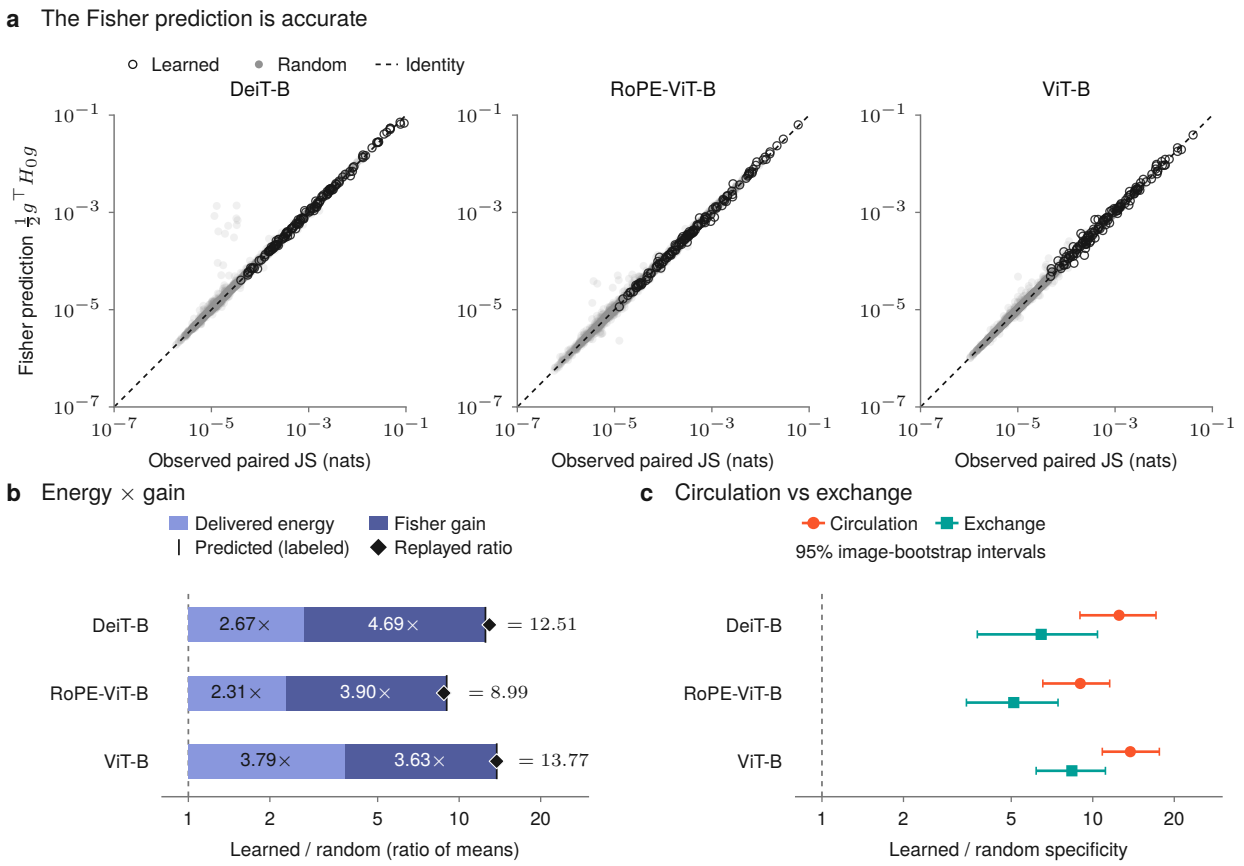


Figure 10.5: **Why learned circulation is consequential** (layer 6, half dose, 128 fresh images per model, 20 feasible alternatives per image and family). **a** The label-free Fisher prediction against replayed paired JS for every learned (dark open circles) and random (gray) direction, with one facet per model. The dashed line is identity. **b** The learned/random ratio of means factors exactly into delivered energy times Fisher gain; diamonds show the replayed ratio. **c** Specificity of learned over random structure for circulation and for exchange at identical canonical norm, with 95% image-bootstrap intervals.

well ($R^2 > 0.997$ over 10,240 responses per model). That is expected of a derivative and does not count against the learned result. The approximation is local, and pooled R^2 hides per-image error: at layer 6 and $a = \frac{1}{2}$ the median relative error of the predicted odd loss is 8–16%, and 42–63% of images are off by more than 10%. It can also fail badly on individual inputs. In an earlier 32-image diagnostic bank, one ViT input carried 99.8% of the squared prediction error, with observed odd loss -0.370 against a predicted -0.044. Halving the dose cuts the root-mean-square prediction error by factors of 7.8–8.0 at the final layer , close to the factor of 8 that the $O(a^3)$ remainder implies, but only by 5.0–6.4 at layer 6, where higher-order terms still contribute at these doses. Descriptively, at layer 6 learned paired JS has Spearman association 0.78–0.86 with the absolute gradient alignment $|\langle \nabla_Z \ell, B \rangle|$, but only 0.07–0.20 with the all-token delivered norm; at the final layer the class-token row of the delivered update is what matters, since the classifier reads nothing else (Table 10.6). We also ran an all-layer analysis on 512 further images (6,144 image–layer cases per model). It includes affine edits at layers whose stationary law was not certified, so we use it only to compare

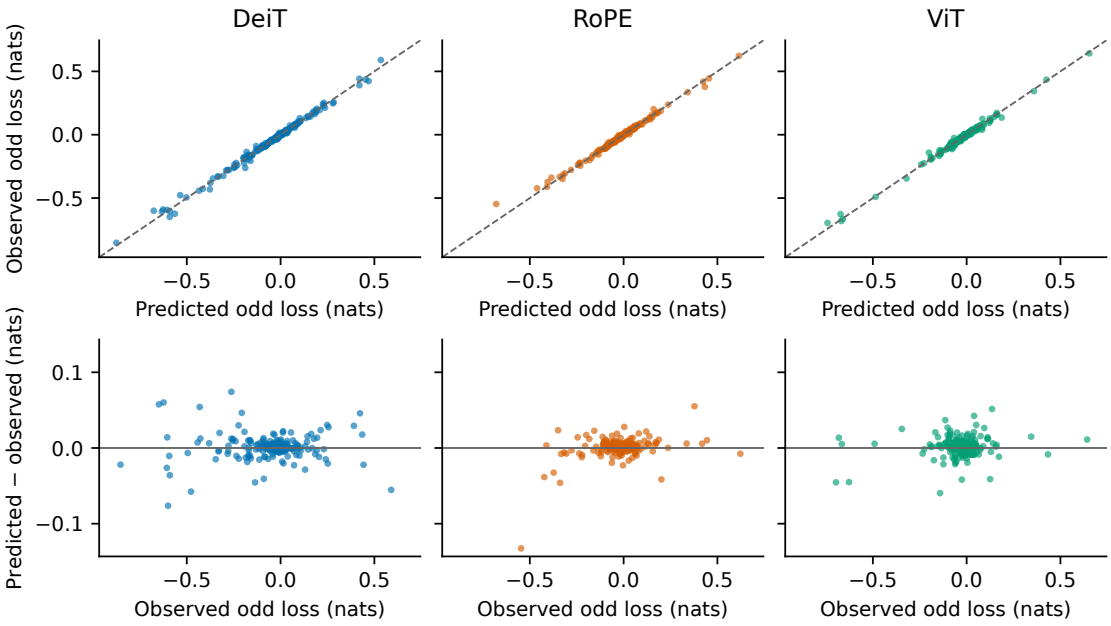


Figure 10.6: **Unfitted prediction of the signed response** (confirmatory study, layer 6, $a = \frac{1}{2}$, 512 images per model). Top: observed odd loss against $a\langle \nabla_Z \ell(Z_S), B_J \rangle$, with the identity line; no slope or intercept is fitted. Bottom: signed residuals. Every observation is shown.

Table 10.6: Descriptive geometry at half current (confirmatory study). Correlations are Spearman associations across 512 learned directions: JS versus delivered norm, absolute loss-gradient alignment, or signed mean row-entropy difference. Cancellation is $1 - \|\sum_h B_h\|_F / \sum_h \|B_h\|_F$, with learned/random means shown. R_E is the learned/random ratio of mean individual JS/$(4\|B\|_F^2)$ in the stated row scope. This is a descriptive normalization, not an equal-delivered-norm intervention.

Checkpoint	Layer	Rows	ρ_{norm}	$\rho_{\|\text{align}\|}$	ρ_{entropy}	Cancellation	R_E
DeiT	6	All	0.203	0.841	-0.121	0.680/0.684	4.36
DeiT	11	CLS	0.772	0.823	-0.016	0.615/0.648	0.93
RoPE	6	All	0.067	0.858	-0.011	0.684/0.689	3.49
RoPE	11	CLS	0.553	0.794	+0.128	0.645/0.656	1.26
ViT	6	All	0.083	0.775	+0.028	0.675/0.686	3.41
ViT	11	CLS	0.612	0.790	-0.223	0.619/0.649	1.16

kinds of information. There, a ridge predictor fitted with image-grouped cross-validation reaches only $R^2 = 0.04$–0.42. It is given every magnitude feature: current, delivered and gradient norms, their products, confidence and operator summaries, but not the sign of the alignment. The unfitted signed rule reaches 0.92–0.99; on an earlier 256-image development bank its R^2 at half dose rises with depth, from 0.81–0.98 at layer 0 to at least 0.9998 at layer 11. *Direction*, not size, carries the prediction.

A consequential orientation need not be a beneficial one. A large paired divergence says that orientation matters. It does not say which orientation helps. On the same 512 images, the learned orientation at layer 6 is loss-favorable on average in DeiT and ViT, with mean odd losses of -0.023 and -0.011 nats whose exploratory 95% bootstrap intervals exclude zero. In RoPE the mean,

Table 10.7: Post-hoc orientation summaries on 256 development or 1,024 confirmation images. Sign counts give images preferring positive/reversed half-current; no ties occur. Divergences and losses are in nats. The DeiT layer-6 confirmation cell, which failed its identity-logit check, is omitted.

Images	Checkpoint	Layer	Median odd	+/− count	Median $JS_{1/2}$	Mean JS_1
Development	DeiT	6	-0.00320	176/80	0.00079	0.01479
Development	DeiT	11	+0.00206	105/151	0.00032	0.00478
Development	RoPE	6	+0.00123	109/147	0.00024	0.00854
Development	RoPE	11	+0.00767	84/172	0.00123	0.01053
Development	ViT	6	+0.00059	122/134	0.00048	0.00918
Development	ViT	11	+0.00652	71/185	0.00052	0.00414
Confirmation	DeiT	11	+0.00254	409/615	0.00033	0.00653
Confirmation	RoPE	6	+0.00088	455/569	0.00027	0.01058
Confirmation	RoPE	11	+0.00666	381/643	0.00133	0.01341
Confirmation	ViT	6	-0.00006	514/510	0.00052	0.01105
Confirmation	ViT	11	+0.00614	314/710	0.00057	0.00430

-0.005 nats, has an interval that includes zero ($[-0.012, 0.002]$), and 276 of 512 images favor the reversed orientation. The sign varies from image to image in every model. At the final layer the picture inverts: *reversing* half the learned current lowers the true-label loss on 310, 347 and 383 of 512 images. Mean odd losses there are -0.007, $+0.003$ and $+0.005$ nats: a minority of large effects favor the learned orientation in DeiT, while ViT's mean favors the reversed one. Accuracy barely moves. Full reversal at layer 11 changes 20, 23 and 11 predictions with net changes of $+1$, 0 and $+1$ correct. At layer 6, 46, 25 and 36 changed predictions yield net changes of -2, 0 and -12. Proposition 10.5(c) explains why the final layer is special. The classifier reads only the class token, and reversal exchanges what that token reads with what reads it. Learned circulation is therefore consequential and aimed, but it is not a loss-minimizing direction, and the model is not at an optimum of the orientation. Two independent orientation studies, on a development bank of 256 and a confirmation bank of 1,024 further images, replayed $\alpha \in \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$ at layers 6 and 11 and reproduce these patterns. The final-layer majority preference for the reversed orientation appears in both banks: 615, 643 and 710 of 1,024 confirmation images favor reversal at layer 11 (Table 10.7). One cell, DeiT layer 6 on the confirmation bank, failed its identity-logit check and is excluded; it is not a scientific null. An earlier control compared the learned current with individual feasible currents. Its loss displacement from the stilled midpoint has RMS 2.4–3.5 times that of individual random draws at half dose. The learned loss lies outside all twenty draws on 49–64% of images, against 9.5% expected under exchangeability, and it occupies both tails. Averaged losses of learned and random currents agree to within ± 0.05 nats: learned circulation is unusual, not uniformly better.

The instrument transfers to a language encoder. The same instrument applies unchanged to masked-token prediction. In RoBERTa-base (Liu et al., 2019), at layers 6 and 11 with all heads intervened jointly, 64 randomly drawn WikiText-2 validation lines (Merity et al., 2017) with the middle token masked all pass the stationary and replay checks. Identity replay reproduces the logits exactly in `float64`. Paired JS is positive on every input at both sites and both doses. The unfitted midpoint gradient predicts the sign of the odd loss on 64 of 64 inputs at each site and dose. The three-gradient Simpson rule of Proposition 10.5(a) reduces the half-dose prediction error by factors of 9.3 (layer 6, RMSE 0.0040 → 0.00043 nats) and 12.4 (layer 11, 0.027 → 0.0021; Table 10.8). The middle-position rule selected six literal `<unk>` tokens among the 64 targets, and all 64 inputs were kept; on the other 58, sign agreement stays at 100% at both sites and doses, with half-dose

Table 10.8: Exploratory masked-token replay in RoBERTa-base, 64 inputs per row. All inputs are retained. Prediction RMSE and odd-loss RMS are in nats; the midpoint predictor uses no fitted coefficient. The Simpson correction is evaluated only at half dose. Sites and doses share the same input sample.

Layer	Dose	Mean JS	Odd RMS	Midpoint RMSE	Sign agreement	Simpson RMSE
6	0.25	0.00154	0.0701	0.000499	64/64	–
6	0.50	0.00608	0.1400	0.003967	64/64	0.000427
11	0.25	0.00604	0.1233	0.003578	64/64	–
11	0.50	0.02205	0.2541	0.026516	64/64	0.002130

Table 10.9: Retrospective classification-error discrimination on the confirmation orientation bank. Scores are paired JS at $a = 1/2$ and $1 - \max_c p_c$ from the identity arm. Higher AUROC is better. Parentheses give pointwise 95% intervals from 2,000 paired image-bootstrap draws. Each row has 1,024 images; the DeiT layer-6 cell remains excluded by its identity-logit check. This reanalysis was not a prespecified endpoint.

Checkpoint	Layer	Errors	JS AUROC	Confidence AUROC
DeiT	11	175	0.790 (0.749, 0.827)	0.839 (0.801, 0.872)
RoPE	11	159	0.709 (0.665, 0.753)	0.868 (0.836, 0.897)
ViT	11	162	0.717 (0.674, 0.758)	0.885 (0.854, 0.911)
RoPE	6	159	0.809 (0.772, 0.843)	0.868 (0.836, 0.897)
ViT	6	162	0.783 (0.747, 0.817)	0.885 (0.854, 0.910)

midpoint and Simpson RMSEs of 0.0034/0.00044 nats at layer 6 and 0.028/0.0022 at layer 11. This is a transfer demonstration on one checkpoint and short inputs, not a language-model benchmark: it demonstrates certified replay, signed prediction and the quadrature refinement in a text encoder, and it does not compare learned with random currents in language models or characterize head functions (Bruce, 2026g, Appendix H).

Paired divergence detects errors less well than confidence. Misclassified images carry a disproportionate share of the orientation response. At layer 6 the roughly 18% of the 256 development images that are misclassified contribute 78–85% of the squared odd response. A natural idea is to use the label-free paired JS to flag errors. On all eleven valid development and confirmation cells, it underperforms the maximum-softmax baseline (Hendrycks & Gimpel, 2017); Table 10.9 lists the five confirmation cells. Paired JS is a sensitivity measure, not a substitute for confidence.

10.5 Circulation implements routing

The vision experiments show that learned circulation matters, but not what it computes. A task with a known answer shows that. In *cyclic retrieval*, 32 tokens form four groups of eight, and each group carries a randomly oriented cycle defined by content. Every token holds a random address code, a pointer equal to its successor's address and a Gaussian payload. Token order is shuffled, so position carries no information about the cycles. *Successor* and *predecessor* retrieval ask every token for the payload of its neighbor in one direction. *Aggregation* asks for the mean payload of the other seven group members, a task with no direction. Models have width 128, four heads per layer, pre-LayerNorm residual blocks with MLPs and one, two or four layers. Each configuration is trained with five seeds, and the train and test pools of codes are disjoint (Bruce, 2026g, Appendix G).

- *Reversing circulation reverses the computation.* Standard attention solves both directional tasks at every depth, already with one layer: about 100% retrieval and payload $R^2 \geq 0.9997$ for every seed. It does so with maximal circulation. The mean over five seeds of the first-layer median circulation index is 1.000. In two-layer successor-trained models those heads send 95% of their attention to the successor and 0.2% to the predecessor, and second-layer heads have $\kappa_{\mathrm{circ}} = 0.001$; later layers of four-layer models stay below 0.01. Replaying the first layer along the matched path (Figure 10.7a) redirects successor-trained models to their predecessors under full reversal (100.0% of queries), and predecessor-trained models to their successors. In the two-layer successor-trained models, half doses retrieve the successor at $\alpha = \frac{1}{2}$ and the predecessor at $\alpha = -\frac{1}{2}$, each about 98% of the time. Stilling produces the nearly equal mixture expected from S. The payload coefficients are 0.495/0.495, and 23% of outputs land nearest the successor, 23% nearest the predecessor and 42% nearest a payload outside the group, because the average of two independent eight-dimensional payloads is often closer to a third. For an ideal permutation head the payload R^2 along the path is $1 - (1-\alpha)^2/2$. The trained models give 1.000, 0.872, 0.493, -0.136 and -1.024 at $\alpha = 1, \frac{1}{2}, 0, -\frac{1}{2}, -1$, against 1, 0.875, 0.5, -0.125 and -1. Reversal at any later layer has no effect: in the two- and four-layer models, retrieval stays at 100% whichever later layer is reversed. These numbers reproduce, with an independent implementation, those of an earlier study (predecessor retrieval 99.91% and 99.84% after reversal; stilled mixture 23.4/23.6/42.5%). The cycle bound of Theorem 10.1(a) holds in all 71,680 head–cycle pairs of the standard models. In the two-layer successor models, first-layer cycle affinities (median 17.0 per token) exceed the successor log-odds they pay for (median 9.4), as the bound requires.

- *Without circulation, direction must be computed.* Theorem 10.1 forbids a circulation-free head from routing around a cycle, and three exactly circulation-free arms obey it to machine precision: symmetrized scores $\frac{1}{2}(QK^\top + KQ^\top)$, symmetrized scores with a learned key bias (sinks allowed, cycles not) and tied $Q = K$. In all 230,400 head–cycle pairs of these models, covering every head of every layer, the geometric mean of successor-to-predecessor odds around the group cycle equals one to within 5×10^{-15}. Yet networks built from symmetrized heads still solve directional retrieval (Figure 10.7b, Table 10.10). They retrieve the successor 96.1%, 98.2% and 98.5% of the time with one, two and four layers (95.4%, 97.8% and 98.2% with a key bias), and the predecessor equally well (94.3–98.6%). They pay in fidelity, with payload R^2 of 0.92–0.97 against 0.9999. Their first-layer heads exploit a freedom that Theorem 10.1(a) leaves open: it fixes geometric means around a cycle, not arithmetic ones. The heads put 46–51% of their attention on the neighbor the task asks for and 32–37% on the other neighbor, although the geometric means of the two are exactly equal on every cycle (Table 10.11; predecessor-trained models are the mirror image). They achieve this by reading unevenly. In successor-trained models, the geometric mean of successor probabilities around a cycle averages 0.05–0.10 and stays below 0.47, under the theorem's cap of one half, while the arithmetic mean is 0.50. Detailed balance ties each predecessor weight to a successor weight, $q_{t+1} = (\pi_{i_t}/\pi_{i_{t+1}})\, p_t$, so the imbalance is carried by preference: the heads read forward most strongly where the successor holds more stationary mass than the reader (mean successor probability 0.72–0.75 on such edges, against 0.21–0.25 elsewhere). Each token therefore receives an unequal mixture of its two neighbors, with weights that vary from token to token and from head to head. A token knows its own pointer, so nonlinear processing of several such mixtures can separate the successor's payload from the predecessor's. Networks of reversible heads thus *compute* direction rather than route it. The complete trained mechanism has not been identified.

- *Tying queries to keys removes more than circulation.* Tied $Q = K$ lags far behind. With one layer, tied models behave like a standard model whose circulation has been stilled. Their outputs

weight the successor's and the predecessor's payloads equally (coefficients 0.36/0.36, against 0.495/0.495 for stilled standard models), and they retrieve the successor only 22% of the time (stilled standard models: 23%). With two and four layers they reach 58–60% and 76–77%, with payload R^2 below 0.7. Theorem 10.1(e) explains the difference. Tied scores are a Gram matrix, so on geometric average around a cycle a tied head cannot attend to either neighbor more than to the reading token itself. The trained heads sit well inside this limit. Summed around each group cycle, $\log(P_{\text{succ}}/P_{\text{self}})$ is at most -10.4, -2.2 and -3.1 in one-, two- and four-layer models, and first-layer heads spend 69–70%, 31–32% and 31% of their attention on the reading token. These sums are computed from the captured probabilities; computed instead from the scores, in which the row normalizers cancel, they agree to within 6×10^{-14}. Approaching the limit would require nearly equal queries for consecutive tokens (10.1), which would make the whole group look alike to the head. Symmetrized heads, whose form is indefinite, prefer both neighbors to the reading token on at least 99.9% of cycles and give it only 2% of their attention. Tying thus removes, together with circulation, the preference for neighbors over the reading token that these symmetrized models use. This reconciles two findings in the literature on sharing query and key projections: tying $Q = K$ is reported to break directionality (Kayyam et al., 2026), whereas shared projections with diagonal rescaling, whose score form is symmetric and can be indefinite, train competitive encoders (Kowsher et al., 2025). Both designs are circulation-free (Table 3.1), but only the indefinite one can prefer a token's neighbors to the token itself. The restriction is structural and holds at any width. Tied models also have half as many query–key parameters as the other arms, but doubling their query width to remove that difference leaves one-layer retrieval at 23% and raises two-layer retrieval only to 66% (Table 10.10).

- *Circulation appears where direction is needed.* Aggregation, which needs no direction, is learned by every arm at every depth (mean retrieval 98–100%; Figure 10.7c), and best by tied $Q = K$ (100%, $R^2 \geq 0.99$), whose symmetric similarity suits the task. Aggregation-trained models of every arm read the whole group almost uniformly (11–12% on each member, including the reading token). Standard attention trained on aggregation learns nearly circulation-free first-layer heads: the mean across five seeds of their median circulation index is below 0.001. The same architecture trained on successor retrieval sits at 1.000. Gradient descent spends circulation only where the task requires direction, and there it uses circulation to route in a single hop.

10.6 Reversible mixers composed in depth

Algebra settles whether composing reversible heads can create circulation. If two heads A and B are reversible with respect to one strictly positive law π, their composition AB keeps π, its reversal is BA, and its circulation is

$$J(AB) = \tfrac{1}{2}\,(AB - BA),$$

so the composition circulates exactly when the two heads do not commute, and AB has the full orthogonal split of Section 3.2. Two reflections of three tokens under the uniform law compose to the cyclic shift, the extreme case, as two reflections of the plane compose to a rotation. The identity is checked in Lean. It assumes one common law, which consecutive layers of a transformer need not share; whether such a composition routes a cyclic successor relation with a linear readout is the test below.

Pending: Whether two reversible mixers composed without an intervening nonlinearity route a cyclic successor relation with a linear readout, which a single reversible head cannot route (Section 10.1), and whether the composition needs a shared cyclic coordinate.

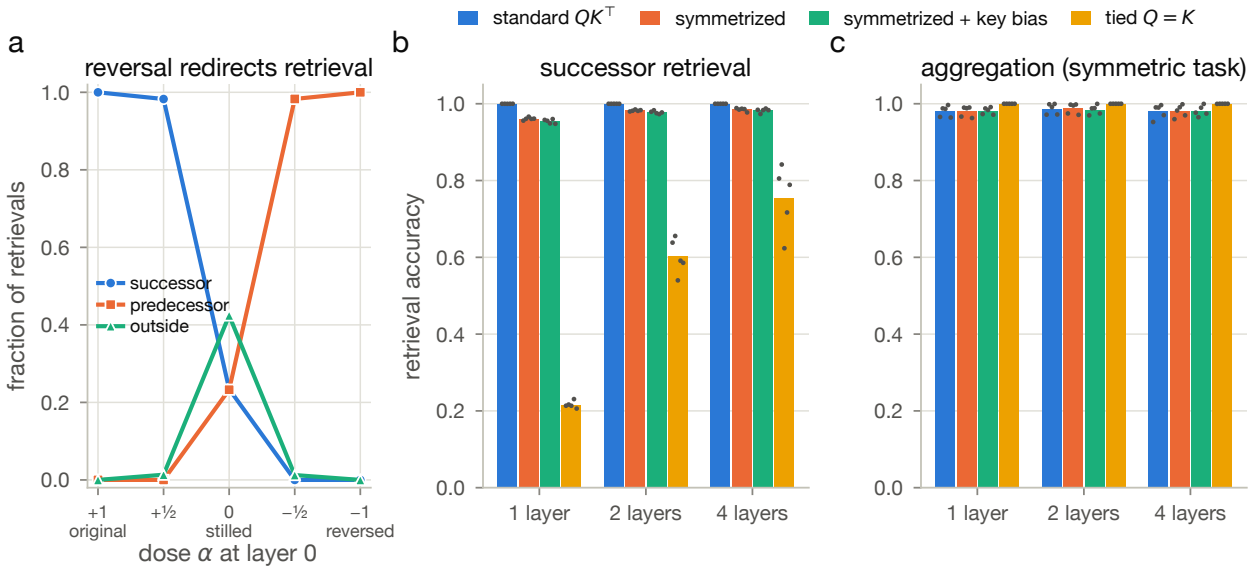


Figure 10.7: **Circulation implements routing.** **a** Two-layer standard models trained for successor retrieval, with all first-layer heads replayed along the matched path. The curves show where the output lands: at the successor's payload, at the predecessor's, or outside the group. Means over five seeds. **b** Successor retrieval by attention arm and depth; bars are means and dots are seeds. The three non-standard arms are exactly circulation-free. **c** Aggregation, a task without direction.

Table 10.10: **Cyclic retrieval grid.** Test retrieval accuracy in percent (and payload R^2), mean over five seeds, by task, attention arm and depth. Last column: mean over five seeds of the median circulation index κ_{circ} of the first layer's heads in each two-layer model. All arms except the standard one are exactly circulation-free.

Task	Arm	1 layer	2 layers	4 layers	First-layer κ_{circ}
successor	standard	100.0 (1.000)	100.0 (1.000)	100.0 (1.000)	1.000
	symmetrized	96.1 (0.944)	98.2 (0.969)	98.5 (0.973)	0.000
	symm. + key bias	95.4 (0.937)	97.8 (0.963)	98.2 (0.970)	0.000
	tied $Q = K$	21.6 (0.322)	60.2 (0.590)	75.5 (0.669)	0.000
	tied, 2× width	22.7 (0.346)	65.6 (0.635)	–	0.000
predecessor	standard	100.0 (1.000)	100.0 (1.000)	100.0 (1.000)	1.000
	symmetrized	95.9 (0.942)	97.7 (0.964)	98.4 (0.972)	0.000
	symm. + key bias	94.3 (0.924)	97.6 (0.963)	98.6 (0.974)	0.000
	tied $Q = K$	21.8 (0.323)	57.6 (0.575)	77.0 (0.685)	0.000
aggregation	standard	98.0 (0.902)	98.7 (0.933)	98.0 (0.921)	0.000
	symmetrized	98.0 (0.903)	98.7 (0.937)	98.0 (0.924)	0.000
	symm. + key bias	98.1 (0.909)	98.4 (0.931)	98.1 (0.923)	0.000
	tied $Q = K$	100.0 (0.996)	100.0 (0.994)	100.0 (0.993)	0.000

Table 10.11: **How first-layer heads read the cycle** (successor retrieval). Attention to the reading token, its successor and its predecessor, in percent, averaged over heads, tokens and five seeds on 256 fresh test sequences. Around each group cycle and in each head, GM(p) is the geometric mean of the successor probabilities p_t, reported as its mean and maximum over head–cycle pairs; Theorem 10.1(a) caps it at $\frac{1}{2}$ for circulation-free heads. The last two columns evaluate (10.1): the share of head–cycle pairs with $\sum_t \log(p_t/P_{i_t i_t}) > 0$, and the largest value of that sum, which Theorem 10.1(e) requires to be nonpositive for tied heads. Self-preference uses 64 fresh test sequences.

		Attention (%)			GM(p)		Self-preference	
Arm	Depth	self	succ.	pred.	mean	max	succ. > self (%)	max $\sum_t \log(p_t/P_{i_t i_t})$
standard	1	0.1	98.3	0.1	0.982	1.000	100.0	+284.9
	2	0.3	95.1	0.2	0.951	1.000	100.0	+283.5
	4	0.0	99.4	0.0	0.993	1.000	100.0	+279.6
symmetrized	1	1.8	50.6	36.2	0.095	0.456	100.0	+202.4
	2	2.3	50.7	32.5	0.056	0.423	100.0	+286.4
	4	2.2	50.6	31.7	0.048	0.423	100.0	+284.6
symm. + key bias	1	1.7	50.4	36.3	0.096	0.469	100.0	+220.8
	2	2.0	50.2	34.0	0.065	0.435	100.0	+272.7
	4	2.0	49.9	33.4	0.074	0.439	100.0	+282.5
tied $Q = K$	1	69.7	9.4	9.4	0.066	0.133	0.0	−10.4
	2	31.8	15.8	14.0	0.064	0.162	0.0	−2.2
	4	30.8	14.4	14.3	0.107	0.175	0.0	−3.4
tied, 2× width	1	68.3	10.2	10.1	0.073	0.134	0.0	−11.1
	2	35.7	16.6	13.3	0.058	0.162	0.0	−2.8

Pending: The outcome of that test, with the qualifier on whether a shared coordinate is necessary read separately for each arm.

10.7 The price of order

A machine that reads a store written over time often needs the latest write, not just a matching one. That needs an arrow: a way to tell later from earlier. *Order blindness.* A soft router's weights are functions of its scores. Two identical writes get identical scores, so without a positional term the router splits its mass evenly between them and cannot tell which came first. A hard router that reads only the scores picks one of the two arbitrarily. *Hard arrows are free.* A machine that compares timestamps selects the later write, and one that overwrites on write keeps only the latest. Neither adds anything to the scores, so neither spends any of the router's margin (Proposition 10.6(d)). They pay elsewhere. Timestamps cost $N \ln N \geq \ln N!$ nats of storage, and overwrite answers only latest-value queries, so it gives up the history (Proposition 9.12(e), Remark 9.13).

Soft arrows are priced. A soft router can order writes only through a positional term added to its logits. Fix a query. Key $j \in \{1, \dots, n\}$ receives the logit $\tau c_j + b_j$, where c_j is its content score and b_j is a positional term: any function of the key's position for this query, such as a recency bias, a linear distance penalty or the positional part of a learned score. Let $B = \max_j b_j - \min_j b_j$ be its range. Two identical writes ($c_j = c_k$) at positions $j < k$ are *ordered with ratio* $r > 1$ when the later one receives at least r times the weight of the earlier one. The two weights share one normalizer, so this holds if and only if $b_k - b_j \geq \ln r$, whatever the other keys are. Write $t(n, h)$ for the number of

edges of the Turán graph (Turán, 1941), the complete h-partite graph on n vertices with parts as equal as possible. Proposition 10.6 prices order in three parts: range, shape and, in mean field, the margin above the router's gain wall (Section 9.3).

Proposition 10.6 (The price of order)**.** *Let* $K = \lfloor B/\ln r \rfloor$.

(a) At most $t(n, K+1)$ *pairs are ordered with ratio* r*. A staircase attains this: split the positions into* $K+1$ *consecutive blocks, as equal in size as possible, and set* $b_j = k \ln r$ *on block* k*. As* $n \to \infty$ *the ordered fraction is at most* $K/(K+1)$*, so ordering a fraction* ϕ *of the pairs needs* $B \geq \lceil \phi/(1-\phi) \rceil \ln r$*.*

(b) A linear term $b = \lambda s$*, with positions* s *spread uniformly over* $[0,1]$*, orders a fraction* $(1 - \ln r/\lambda)_+^2$ *of the pairs in the continuum limit. At a large range* $B = \lambda$ *it leaves about* $2 \ln r/B$ *of the pairs unordered, twice the staircase's* $\ln r/B$*.*

(c) (Mean field.) Let the N *background scores be exchangeable with* $m(\tau) = \mathbb{E}e^{\tau c}$*, let the target's score be* c_t*, and require every position to stay retrievable: a single target at the position with the smallest positional term keeps at least half the mass. This costs* $\frac{1}{N}\sum_j e^{b_j - \min b} \leq e^{\Delta}$*, where* $\Delta = \tau c_t - \ln m(\tau) - \ln N$ *is the unbiased router's margin above its gain wall. Over all positional terms that meet it, the largest ordered fraction tends, as* $N \to \infty$*, to*

$$F^*(\Delta) = \max\Big\{1 - \sum_{k \geq 0} f_k^2 \;:\; f_k \geq 0,\ \sum_k f_k = 1,\ \sum_k f_k r^k \leq e^{\Delta}\Big\},$$

attained by a staircase whose block k *holds a fraction* f_k *of the positions. The maximizer has the form* $f_k = (\mu - \nu r^k)_+$ *with* $\nu \geq 0$*. At the gain wall* ($\Delta = 0$) *nothing can be ordered, and* $F^*(\Delta) < 1$ *for every finite* Δ*.*

(d) A hard machine that compares timestamps, or overwrites on write, orders every pair and spends none of this margin.

The proof is in Section A.11. On constructed keys the price is concrete.

Remark 10.7 (The constructed arrow test). The constructed arrow test has $N = 1024$ keys in $D = 128$ dimensions, exact targets $c_t = 1$, background cosines of uniform random unit vectors, and $r = 9$: the later copy must hold 90% of the pair's weight. There, $\Delta = 2.68$ nats at $\tau = 10$ and 11.52 at $\tau = 20$, and Table 10.12 gives $F^* = 0.60$ and 0.86. So, in mean field, no positional term of any shape reaches 0.9 latest-write success while every position stays retrievable. The optimal staircase orders nearly three times as many pairs as the best linear term at $\tau = 10$. A staircase designed for $\Delta - \frac{1}{2}$ reaches its mean-field value in simulation, with a single target on its lowest step still retrieved in 95.9% of trials, so the bound is attained beyond mean field.

Recency carried in soft weights is unary whatever the positional code: an optimized rotary term never exceeds the staircase's bound (Section 9.5).

Constructed-key records. The arrow test shows each piece (Figure 5.2d). Without a positional term, softmax splits the two copies evenly and fails every latest-write query, and a walker without timestamps returns either copy half the time. A walker that compares timestamps succeeds every time, and an address book that overwrites on write succeeds by construction. A linear recency bias of slope λ gives latest-write success that follows $(1 - \ln 9/\lambda)_+^2$ to within 0.02 up to the slope that maximizes it, and then falls as recent distractors gain. Its best on the tested slopes is 0.51 at $\tau = 10$ ($\lambda = 8$) and 0.75 at $\tau = 20$ ($\lambda = 16$). At those slopes the oldest quarter of single items is retrieved 0.00 and 0.60 of the time. No tested slope reaches 0.9 on latest writes. The recovered panel retains all twelve biases at both gains, timestamp/no-timestamp walkers and overwrite by construction.

Table 10.12: The price of order in the constructed arrow test ($N = 1024$, $D = 128$, $r = 9$). Δ is the unbiased router's margin above its gain wall. $F^*(\Delta)$ is the largest fraction of duplicate pairs that any positional term can order while every position stays retrievable (mean field, Proposition 10.6(c)); steps is the number of steps of the staircase that attains it. The staircase designed for $\Delta - \frac{1}{2}$ was run on the test's substrate (20,000 trials): latest write is the fraction of duplicate pairs in which the later copy holds at least half the mass and at least r times the earlier copy's weight, and lowest step is the fraction of single targets on its lowest step that hold at least half the mass. The best linear term is the largest slope λ within the same budget, and ordered is the fraction of pairs it orders, $(1 - \ln r/\lambda)^2$.

		any positional term		staircase for $\Delta - \frac{1}{2}$			best linear term	
τ	Δ (nats)	$F^*(\Delta)$	steps	$F^*(\Delta - \frac{1}{2})$	latest write	lowest step	λ	ordered
10	2.68	0.60	3	0.55	0.55	1.00	4.1	0.22
20	11.52	0.86	8	0.85	0.85	0.96	14.2	0.71

Its saved F^* and best-linear references are mean-field quantities conditional on retaining every position, not measured maxima of the average-accuracy curve. Unique-retrieval coordinates and latest-write coordinates come from their respective saved records. The saved execution failure and historical non-reproduction remain; publication is admitted as a registered rebaseline (Bruce, 2026e, Appendix C.1).

Physical machines pay in entropy. For a stochastic physical machine whose routing is a Markov kernel with a positive stationary law, the arrow also has an energetic price. Split the stationary flow $F = D_\pi P$ into its symmetric part, the reciprocal traffic A, and its antisymmetric part $D_\pi J$, whose entries are $\pi_i J_{ij} = \frac{1}{2} C_{ij}$ (3.2); this is the π-weighted split of Section 3.1, not the plain symmetrization of P. Assume that every edge with positive flow also has positive reverse flow.

Theorem 10.8 (Entropy production at second order)**.** *With sums over ordered pairs, define the entropy production per step as* $\sigma = \sum_{i,j} F_{ij} \ln(F_{ij}/F_{ji}) = \frac{1}{2} \sum_{i,j} (F_{ij} - F_{ji}) \ln(F_{ij}/F_{ji})$. *If* $|\pi_i J_{ij}| \le \eta A_{ij}$ *with* $\eta < 1$ *on every edge with positive flow, then*

$$\sigma = 2 \sum_{i,j} \frac{(\pi_i J_{ij})^2}{A_{ij}} + 2 \sum_{k \ge 1} \frac{1}{2k+1} \sum_{i,j} \frac{(\pi_i J_{ij})^{2k+2}}{A_{ij}^{2k+1}},$$

and the nonnegative remainder after the first term is at most $\frac{2}{3(1-\eta^2)} \sum_{i,j} (\pi_i J_{ij})^4 / A_{ij}^3$. *In particular* σ *is even under reversing the current at fixed* A.

The proof is in Section A.11. This σ is the entropy production $D_{\mathrm{KL}}(F \| F^\top)$ of Theorem 3.1. Circulation therefore costs entropy at second order in the current, whatever its sign (after Schnakenberg, 1976). A one-way flow with zero reverse flow makes σ infinite in this convention, and the expansion does not apply. This is a statement about a physical machine's Markov dynamics: the physical reading assumes a thermodynamic model and its units, and an attention matrix by itself is not a measured dissipation.

10.8 Decoders: the leak rule

The reversible vocabulary does not apply to full causal softmax attention over more than one token, which has no positive stationary law (Section 2.5). For decoders we use only statements proved for

them: absorption at the first token and the sink-collapse law (Theorem 2.16), the survival reading of rollout (Section 8.3), and the two-sided law (Section 9.1). A decoder layer built with a reversal, J or S from a regularized positive reference law must be masked causally and tested for leakage, since these constructions can mix future tokens into earlier outputs (Section 19.5). Much of what matters in decoders, including in-context learning and long-range retrieval beyond a single head, lies outside the lattice (Chapter 14).

In the language of Chapter 8, a causal mask supplies an absorbing state without any null: decoders are killed by construction (Section 8.2), and their arrow comes from the mask rather than from circulation. The mask decides where a sink can form, not whether one does: trained language models place one on the first token, and every rule that gives attention a null removes it (Section 17.4).

10.9 What holds in trained networks

Exact for every network of the stated class. Direction requires circulation (Section 10.1): an identity for every circulation-free head with a positive stationary law, on cycles whose successor and predecessor probabilities are positive, at any layer, stated per head; compositions of reversible heads can acquire direction, and causal heads take theirs from the mask. The price of order (Section 10.7): a combinatorial bound for any soft router whose positional logit term has range B.

Measured in trained networks. Learned circulation, its census and the consequences of reversing it (Sections 10.3 and 10.4): measured in trained encoders. Qualified baselines and the leak rule (Sections 10.8, 18.5 and 19.5): measured on trained baselines; the leak rule itself (reversal objects of a causal operator are anti-causal) is exact.

Table 10.13 collects the boundaries of each claim that this chapter and Chapter 3 take from Bruce (2026g): algebraic statements are exact under their stated assumptions, and empirical statements concern the models, sites and samples named. Three deserve emphasis. First, the stationary construction needs a positive stationary law and reciprocal support. It covers bidirectional attention broadly, but for causal decoders the full-token circulation is identically zero. Studying their directional computation requires different objects, such as bidirectional sub-blocks or return-map constructions, which are not developed here. Second, the empirical panel contains one pretrained checkpoint per architecture. The image and text samples replicate across inputs, not across training runs, and independently pretrained replicates such as MultiBERTs (Sellam et al., 2022) would show how stable the census is across seeds and over training. Third, specificity is measured against declared feasible samplers, which are not uniform over all admissible currents. The comparison matches operator magnitude, stationary law, reciprocal traffic and within-pair spectra, but not delivered norm. The mechanism analysis (Section 10.4) shows that delivered energy is part of what makes learned circulation special.

Table 10.13: **Claims, evidence and limits.** Algebraic statements are exact under their stated assumptions; empirical statements concern the models, sites and samples named.

Claim	Evidence	Limit
Attention asymmetry = preference $\oplus$ exchange $\oplus$ circulation; sink + cycle in scores	Theorems 3.1 and 3.2; 73 executable checks	Needs a positive stationary law and reciprocal support; causal kernels have $C \equiv 0$
Direction on a cycle requires circulation	Theorem 10.1; exact odds identity in trained reversible models	Per head; depth and values can compute direction without it (Section 10.5)
Stilling and symmetrization are canonical; reversal is an attention call	Theorem 10.3 and Proposition 10.4; FlexAttention check	The matched pair fixes the operator, not downstream activations, which may leave the training distribution
Trained encoders circulate, mid-network and in positional heads	Census of 184,320 certified kernels in five models	One checkpoint per architecture; ImageNet and WikiText inputs
Learned circulation is unusually consequential	Confirmatory study fixed before evaluation, 3 models × 512 images; replication on a fresh bank	One declared sampler family; no independent pretraining runs
It is aimed: energy and Fisher gain	Label-free Fisher decomposition; exchange control	First-order analysis; layer 6 only
The delivered direction predicts the signed response	Unfitted R^2 about 0.99 or higher (vision); 64/64 signs (RoBERTa)	Local approximation: at layer 6 and $a = \frac{1}{2}$ the median per-image error is 8–16% despite pooled R^2 of about 0.99 or higher; target-aware
A consequential orientation need not be a beneficial one	Loss and accuracy on 512–1,024 images per model	Describes these checkpoints; no claim of an optimal orientation
Without circulation, direction must be computed rather than routed; tied $Q = K$ cannot prefer neighbors to self	Theorem 10.1(e); retrieval grid, 4 arms × 3 depths × 3 tasks × 5 seeds, plus a width control; both identities hold in every head–cycle pair	Synthetic one-hop task; small models; mechanism of the computation not fully identified

Chapter 11

A writer: where features are made

The six conditions of Chapters 5–10 protect the averaging that moves information between tokens. A seventh condition joins them: transport alone cannot create features, so every machine needs a writer (Bruce, 2026e). Chapter 0 places it outside the inventory of six (Section 0.3.2), because moving information around is not enough, and in the block of Section 1.1 the MLP is the feature writer. This chapter shows why transport cannot write (Section 11.1), separates invertible, volume-preserving and isometric blocks (Section 11.2), and measures what the tested orthogonal transport costs on retrieval and on classification and what invertibility buys (Sections 11.3 to 11.5); what makes a good writer remains open (Section 11.7).

11.1 Transport cannot write

The first reason is elementary: a mixing step that applies a nonnegative row-stochastic matrix, however the matrix is chosen, forms averages of the rows it is given.

Proposition 11.1 (Transport cannot write)**.** *Let $X_{t+1} = P_t(X_t)X_t$ with every P_t row-stochastic and nonnegative, chosen adaptively from the current state. Every output row lies in the convex hull of the rows of X_0, and $\operatorname{rank} X_t \le \operatorname{rank} X_0$. A common pointwise nonlinear writer can raise the rank and leave the hull.*

The proof is in Section A.12. The statement concerns one shared left transport. Several heads with different channel maps break its premise: from the rank-one rows $(1, 0)$ and $(0, 0)$, an identity head keeping the first channel and a swap head writing it to the second sum to the rank-two identity. A shared linear channel map alone still cannot raise the rank. None of this sets the rank a task requires, or gives an accuracy bound without a specified readout (Bruce, 2026e).

Transport that is not an average meets the same limit through an invariant.

Remark 11.2 (Write capacity). Transport that mixes the tokens of a group by one shared orthogonal matrix preserves the group's channel Gram matrix $U^\top U$, so a readout that depends only on that matrix cannot change through the stack; individual token rows still can. Dictionary retrieval R^2 is 0.18–0.26 against at least 0.99996 for additive blocks, and CIFAR-100 accuracy is 10 points lower, measured on trained models (Sections 11.3 and 11.4).

Stated as a test, the writer is the seventh of the conditions that the catalogue of routing machines applies to every machine (Section 19.1) (Bruce, 2026e). Its removal signature, a prediction that is not yet tested, is that no new features appear: nonlinear readouts such as XOR fail. What to measure is accuracy on feature-creation tasks. A machine pays for the writer in MLPs, edge messages, lookup tables, nonlinear devices or reactions, and Proposition 11.1 prices the payment.

11.2 Invertible, volume-preserving and isometric blocks

Three properties of a block are easily run together. A block is *invertible* when its input can be recovered exactly from its output, *volume-preserving* when it keeps volumes, and *isometric* when it keeps distances between inputs; a block that keeps the length of each state is *norm-preserving*, which for a fixed linear map is the same as isometric but not in general. The additive coupling $u' = u + v$, $v' = v + u'$ is invertible, since the two subtractions $v = v' - u'$ and $u = u' - v$ recover every input, and it keeps area, mapping the unit circle to an ellipse of equal area, but norms can change under it. A fixed orthogonal rotation keeps area and norm, and its transpose is its inverse (Figure 11.1a,b). Invertibility therefore permits feature writing without requiring norm preservation. The transport studied in this chapter is conditional orthogonal transport (COT), whose core keeps each state's norm but need not preserve distances between different inputs. Its arms form a ladder: the token-rotation arm, COT-P, preserves group Gram matrices; the channel-rotation arm, COT-CR, permits channel rotations and keeps only each state's norm; and COT-W adds reversible feature writes and does not keep the norm. The additive reversible arm, B-REVADD, and the transformer reference, B-TF, complete the comparison.

Beneath the comparisons lies a layer of exactness checks, each a machine-precision check of an identity. In float64 the three transport stacks invert through 48 blocks to a relative error of 2.4–3.0 $\times 10^{-15}$, at both controller gains tested, and in float32 to 3.1–4.5 $\times 10^{-6}$. In float64 the two norm-preserving stacks hold the state's energy fixed through those 48 blocks to 1.5–2.0 $\times 10^{-15}$, so they neither contract nor amplify at any depth tested; the token-rotation stack keeps its group Gram matrices to 2.0 $\times 10^{-15}$, the channel-rotation stack changes them by a relative 0.0154–0.0155 while keeping the norm, and the stack with feature writes changes the norm. Gradients computed by reconstruction agree with ordinary backpropagation to within a relative 1.5 $\times 10^{-14}$ in float64 at depths 8 and 24, for the additive reversible arm and the three transport arms. The routing comparison uses three seeded synthetic constructions; its rank ratios are seed means. A conservative routing head, the orthogonal routing of Section 5.6, is orthogonal to 1.5 $\times 10^{-15}$, so its condition number is 1.0000, and over 16 layers on 48 tokens it conserves the effective rank of the content it moves (ratio 1.00), where routing the content through the averaging grouping instead collapses it (ratio 0.15). The correctness battery passes all 210 of its checks in twelve suites, in float32 and float64, but its two reference-parity suites could not run, because the reference implementation they compare against was absent, and two numerical-threshold checks, both for one causal arm outside this family, failed in float32; the qualification of these implementations is therefore not complete.

Figure 11.1 sets the two kinds of invertible block side by side, the additive coupling that writes and the rotation that keeps norms, and then the recorded ladder. On it, five seeds each, CIFAR-100 accuracy and single-value retrieval separate classification from retrieval across the transformer reference, the additive reversible arm and three orthogonal-transport arms: the additive reversible arm keeps both, the two norm-preserving arms lose both, and the arm that adds writes recovers classification but not retrieval. These are provisional observations: the qualification of the implementations and matched-parameter controls remain unresolved, and they do not imply that every orthogonal architecture fails retrieval.

11.3 Retrieval: additive coupling solves it, the tested orthogonal transport does not

On the synthetic bidirectional dictionary task the separation is sharp. Standard attention and additive reversible attention keep near-perfect retrieval at every dictionary size tested, from 8 to 64

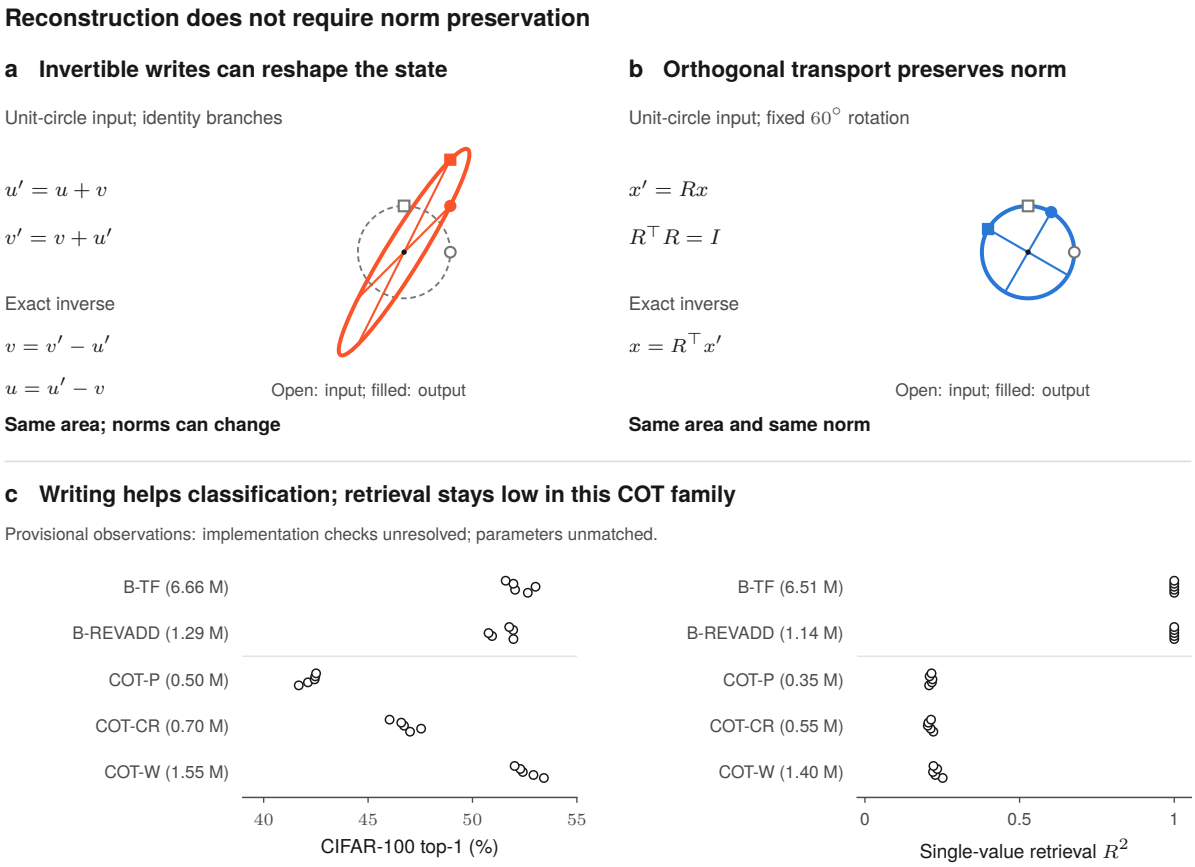


Figure 11.1: **Invertibility permits feature writing without requiring norm preservation, and the recorded ladder of conditional orthogonal-transport (COT) arms separates classification from retrieval.** (a) Exact illustrative construction: the additive coupling with identity branches maps the unit circle to an ellipse of equal area; the two subtraction steps recover every input. (b) A fixed 60° orthogonal rotation preserves the unit circle and has its transpose as inverse. Both diagrams use the same coordinate scale. Open and filled circle/square markers identify corresponding input and output vectors; solid diameters are the images of the input coordinate diameters. The conditional COT core preserves each state's norm but need not preserve distances between different inputs. (c) Each open mark is one of five seeds (11, 23, 37, 53, 71): CIFAR-100 test top-1 on 10,000 images (structural arms with query-pool readout), and single-value retrieval R^2 at 16 associations on 2,048 examples. Vertical offsets separate seeds; seed-to-seed variation is shown directly, with no confidence interval. B-TF is the transformer reference and B-REVADD the additive reversible arm; COT-P preserves group Gram matrices, COT-CR permits channel rotations, and COT-W adds reversible feature writes. Parentheses give parameter counts in millions. These are provisional observations: the qualification of the implementations and matched-parameter controls remain unresolved. They do not imply that all orthogonal architectures fail retrieval.

associations, both for single values and for differences of two values, with seed-mean R^2 at least 0.99996, while the input-conditioned COT arms stay at seed-mean R^2 0.18–0.26 at the training size of 16 associations, and every tested COT arm falls below 0.03 at 64 (Figure 11.2). The rotation controls do not move it. Angles fixed at zero or learned independently of the input give R^2 near zero; self-conditioning, a larger initial gain and a wider controller stay with the other rotation arms; and the arm with reversible feature writes fails with more parameters than the additive reversible arm. The cells are provisional and concern this family only.

11.4 Classification: writing restores parity

Classification tells a different story. Adding a feature-writing channel to the transport restores parity on CIFAR-100 classification but not on dictionary retrieval, where mean R^2 over five seeds stays at 0.19–0.23 across single-value and difference retrieval at the training size of 16 associations, measured on trained models (Bruce, 2026d). In the recorded comparison (structural arms with a query-pool readout, five seeds; Figure 11.1c), the transformer reference reaches 52.25% top-1 and the arm with writes 52.61%, higher in four of the five paired seeds and by 0.36 points on average; the additive reversible arm sits 0.78 points below the reference (paired difference, 95% interval $[-1.98, 0.43]$ over the five seeds) with a fifth of its parameters, and the token-rotation and channel-rotation arms lose 10.01 and 5.47 points. The arms are not parameter-matched: they range from 0.50 to 6.66 million parameters.

Why transport cannot write is the subject of Section 11.1: one shared averaging transport keeps every output in the hull of its inputs and cannot raise their rank , and transport that preserves the Gram matrix fixes every readout that depends only on that matrix. The ladder of Section 11.2 releases the Gram invariant and then the norm, one step at a time, and classification follows the release: CIFAR-100 top-1 is 42.24%, 46.78% and 52.61% from the token-rotation arm through the channel-rotation arm to the arm with writes, although the arms also grow from 0.50 to 1.55 million parameters. Retrieval does not follow it: single-value R^2 at 16 associations is 0.21, 0.21 and 0.23 along the same ladder. On this ladder, writing restores classification and leaves retrieval where it was.

11.5 Reconstruction saves memory

Invertibility has a second use, reconstruction: a block's activations are recomputed from its output during the backward pass instead of being stored. On the recorded profiles, reconstruction lowers the allocation increment of a training step, while ordinary autograd has the lowest median step time in every architecture–depth comparison (Figure 11.3). At depth 48 the increment under reconstruction is about 0.24–0.29 of the increment under checkpointing for the four invertible arms, and once one parameter-sized buffer is subtracted it stays nearly flat from depth 8 to 48; autograd remains the fastest at every depth. Reversibility buys memory, not speed. These are implementation observations from one seed, on an unrecorded hardware model, not architectural speed claims, and the accounting does not establish asymptotically constant memory, since parameter and optimizer storage grow with depth.

11.6 Against established reversible and orthogonal designs

Pending: Dictionary-retrieval R^2 and CIFAR-100 top-1 for established reversible and orthogonal designs, measured under the same protocol as the arms of this chapter.

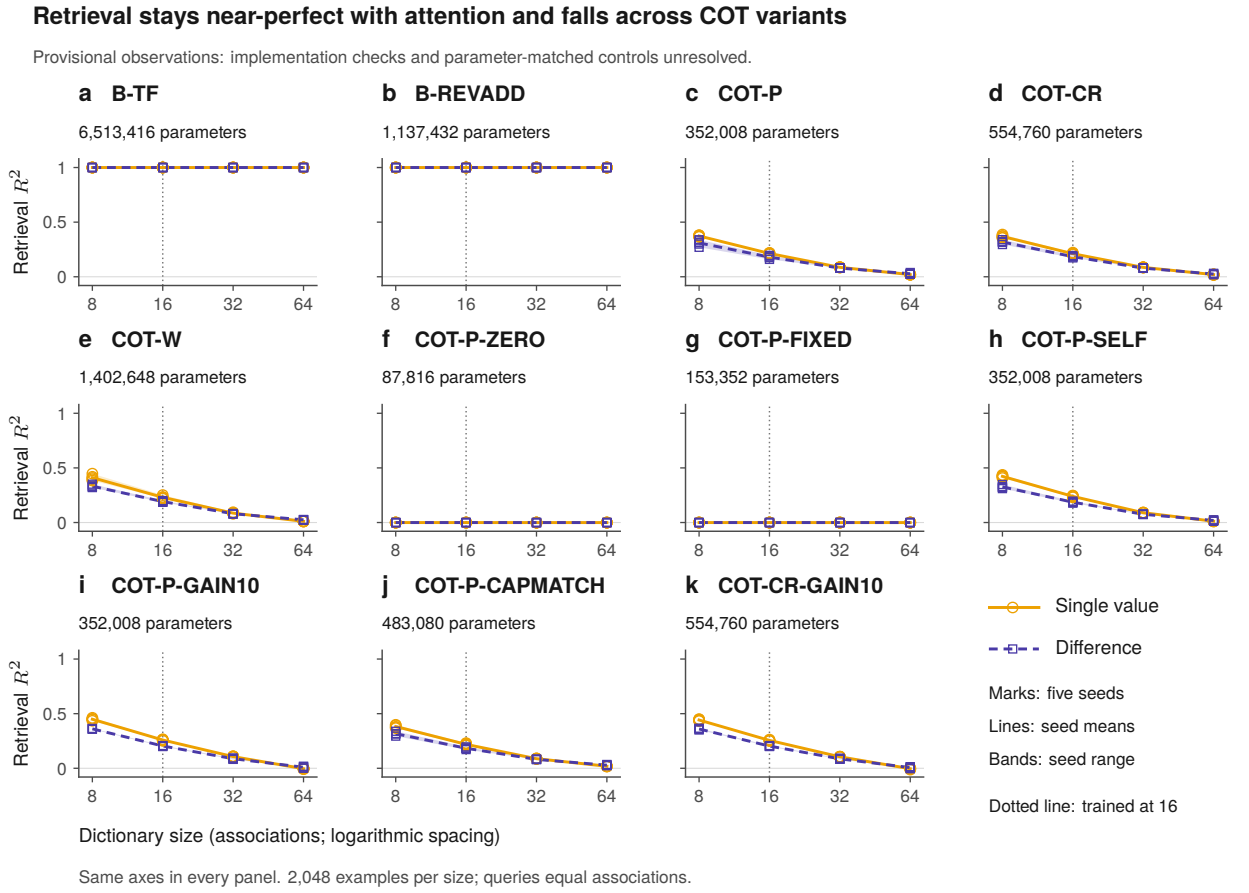


Figure 11.2: **B-TF and additive reversible attention retain near-perfect retrieval across dictionary sizes, while every tested COT arm has seed-mean $R^2 < 0.03$ at 64 associations.** (a–e) The architecture ladder: standard attention, additive reversible attention, Gram-preserving token rotations (P), additional channel rotations (CR), and reversible feature writes (W). (f–k) Rotation controls: ZERO fixes angles at zero; FIXED learns input-independent angles; SELF uses self-conditioning without a guaranteed inverse; GAIN10 starts the effective gain at 0.1 instead of 0.01; CAPMATCH widens the P controller from 64 to 96 units. Its total parameter count is not matched to CR or B-TF. Counts above each panel include all trainable parameters. Circles and solid lines denote single-value retrieval; squares and dashed lines denote difference retrieval, with target $(v_i - v_j)/\sqrt{2}$. Open marks show all five seeds (11, 23, 37, 53, 71); lines join seed means; bands span the seed minimum and maximum, not a confidence interval. Coincident marks remain superposed. All panels share axes, including negative values. Each model trained for 10,000 steps at 16 associations (dotted line); evaluation uses 2,048 examples per size and equally many queries as associations. The recorded score is $R^2 = 1 - \sum \|\hat{y} - y\|^2 / \sum \|y\|^2$, against the zero-mean target population. These 110 cells remain provisional: the qualification of the implementations and matched-parameter controls are unresolved. Larger dictionaries activate pairing stages absent at training length; the observations concern this COT family and do not establish a universal limitation of orthogonal retrieval.

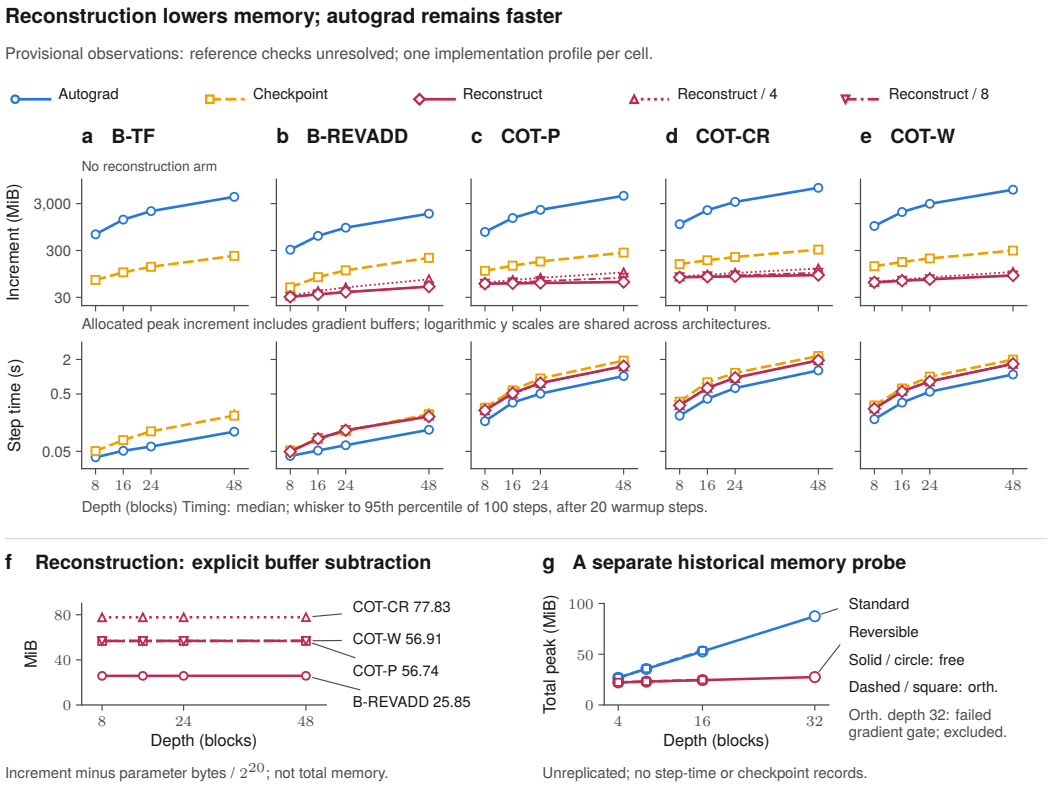


Figure 11.3: **Reconstruction lowers the recorded allocation increment, while ordinary autograd has the lowest median step time in every architecture–depth comparison.** (a–e) Five architectures, with identical logarithmic resource scales across columns. The upper row subtracts the no-gradient forward peak from the forward/backward allocated peak; this increment includes newly allocated parameter gradients. MiB means 2^{20} bytes. Checkpointing uses one non-reentrant checkpoint per block; reconstruction variants /4 and /8 retain segment boundaries every four or eight blocks. Coincident marks overlap. The lower row shows synchronized complete optimizer-step medians and whiskers to the 95th percentile, including recomputation, clipping and optimizer updates: 20 warmup and 100 measured steps per cell. These are within-run timing spreads, not confidence intervals. All 88 recorded cells use seed 0, width 256 and a fixed batch of 64 CIFAR-100 examples; each has one unreplicated memory probe and passes its gradient check. (f) Reconstruction increment minus one parameter-sized buffer, with depth-48 values labelled. This accounting diagnostic does not isolate all activation allocations or establish asymptotic constant memory; parameter and optimizer storage grow with depth. (g) Total allocated peaks from a separate historical memory probe on a routing task: seven valid seed-0 gradient probes, free or orthogonal branches. The orthogonal depth-32 record has null gradient and reconstruction errors and is excluded; its failure is not evidence of norm overflow. The historical probe records no comparable step-time or checkpoint measurement. The qualification of the implementations remains unresolved, and resource profiling alone does not establish that an arm matches the reference in accuracy. The hardware model is unrecorded, so these are implementation observations, not architectural speed claims. The historical probe is unreplicated and its source code is not fully recovered; panels (a–f) do not reproduce it.

11.7 The MLP as the writer and the questions left open

The transport-only statements therefore hold for the pure transport mixer; variants that re-center or write break its invariant. In transformers the MLP and the output map W_O write, and the MLP is the one block component whose theory remains open (Bruce, 2026d). Among the block protections of the lattice (Section 14.1), write capacity is the row whose removal signature is that pure transport cannot write, whose compensation is partial, since a write channel restores classification but not retrieval, and which practice pays for with the MLP and W_O. Chapter 0 reads the usual description of the block, in which attention moves information between tokens and the MLP transforms each token (Section 0.2.2), through this condition: the MLP is the writer (Table 0.1).

The lattice of protections (Chapter 14) treats the MLP as the feature writer and as the main currency of compensation (Section 15.1), and gives no theory of either role. Pure transport cannot write (Remark 11.2). In the tested transport family, a writing channel restores CIFAR-100 classification without restoring dictionary retrieval (Section 11.4), in a comparison whose arms are not yet parameter-matched; this says that writing matters but not what a good writer is. How much of each protection's price the MLP can absorb, and at what width, is the main open problem the framework leaves (Bruce, 2026d).

Pending: Where each routing machine of the catalogue makes its features, and which writer it pays with.

11.8 What holds in trained networks

Exact for every network of the stated class. The limits on transport (Section 11.1): averaging keeps every output in the hull of its inputs and cannot raise their rank, for every row-stochastic mixer; shared orthogonal mixing preserves each group's channel Gram matrix. Both hold at every layer.

Measured in trained networks. Invertible blocks that keep capability, and the tested orthogonal transport that loses it (Sections 11.3 to 11.5): measured on trained mixers; the invariant behind it is exact (shared orthogonal transport preserves each token group's channel Gram matrix), but it does not imply the retrieval failure.

Pending: there is nothing yet to transfer. What makes a good writer (Section 11.7). Open: there is no theory yet to transfer.

An interpretation, with nothing to transfer. The partial coverage of decoders and of the MLP, which the theory treats only as the writer (Section 11.7; Section 0.6.3): a statement of scope.

Part III

When averaging feeds back

Chapter 12

Converged averages can amplify

Part II named what an averaging layer that is rebuilt from its input and iterated needs, and Chapter 7 certified one of those needs, gain control: when the anchor outweighs the feedback, $m = \beta - \alpha\kappa_{\mathrm{f}} > 0$, the anchored iteration (1.4) converges to a unique fixed point (Theorem 7.1). This part asks what a settled average still does. Chapter 0 gives the answers in brief (Section 0.4.3). A settled average can amplify changes in its inputs and overturn the predictions read from it, which is the subject of this chapter. And settling certifies less than it seems: a computation that settles to machine precision certifies neither its response, nor its uniqueness, nor that it attracts, nor that a loop built on it is stable. That is the subject of Chapter 13, where it is proved and shown on constructions and small systems; trained looped and implicit models are still to be tested.

This chapter develops the first answer from its two sources (Bruce, 2026h) and does not repeat Chapter 0's summary. Section 12.1 separates the weights that represent an equilibrium from the derivative that describes its response: every equilibrium is a positive average of the anchors, yet a smooth positive mixer can approach the optimal response bound β/m while its frozen average approaches the identity. Sections 12.2 and 12.3 find amplification in ordinary softmax kernels and in a photographic denoiser. Section 12.4 turns from sensitivity to classification: converged mixing over contexts of 32 images overturns class margins that frozen weights keep. Section 12.5 gives the exact accounting behind it, each prediction's margin split into its own and its companions', and Section 12.6 uses it to take apart a focal failure. Sections 12.7 and 12.8 mark what is still to come, and Section 12.9 says what holds in trained networks. The maxima reported here were selected from grids: they show what can happen, not how often it does.

12.1 Positive averages with a signed, amplified response

Operator norms in this chapter are induced by the maximum-row norm $\|\cdot\|_{\infty,2}$ of Section 1.4 unless stated otherwise. At an equilibrium X^* of (1.4), write $P_* = \mathcal{P}(X^*)$ for the realized weights, and DF for the Jacobian of the update F_A at X^* (Bruce, 2026h).

Proposition 12.1 (The frozen smoother and the input response)**.** *Holding the realized weights fixed gives*

$$X^* = R_* A, \qquad R_* := \beta(I - \alpha P_*)^{-1} = \sum_{k=0}^{\infty} \beta\alpha^k P_*^k. \tag{12.1}$$

The coefficients in this series are nonnegative and sum to one. Consequently R_* is row stochastic and nonexpansive in the maximum-row norm: each output is a convex combination of the anchors.

On a differentiable equilibrium branch with invertible $I - DF$,

$$DF[U] = \underbrace{\alpha P_* U}_{\text{changing values}} + \underbrace{\alpha(D\mathcal{P}(X^*)[U])X^*}_{\text{changing weights}}, \qquad H := D_A X^* = \beta(I - DF)^{-1}. \tag{12.2}$$

The proof is in Section A.13. The response to changing the anchors includes the adaptive-filter product rule emphasized by Reehorst & Schniter (2019). Local attraction, $\rho(DF) < 1$, suffices for that invertibility. Unlike R_*, H need not be nonnegative or nonexpansive. Contraction compares trajectories with the *same* anchors; input response compares equilibria with *different* anchors. Rebuilding can therefore make an increase in one input decrease another output without violating positivity of any averaging weight.

Chapter 7 bounded the response by the margin: on a common invariant ball, two anchors satisfy $\|X^*(A') - X^*(A)\|_{\infty,2} \le (\beta/m)\|A' - A\|_{\infty,2}$ (Corollary 7.2). Positive mixers whose frozen average approaches the identity can approach that bound (Bruce, 2026h).

Proposition 12.2 (Optimal response from positive mixers)**.** *For every $0 < \alpha < 1$ and $0 < m < \beta$, there are smooth two-token builders contractive on a common unit ball with upper factor $1 - m$, and interior anchors, such that $R_* \to I$ and $\|D_A X^*\| \to \beta/m$ (in both induced maximum-row and Euclidean norms). Cross-token derivatives are negative sufficiently far along the family. With $\alpha = \beta = 1/2$, a further family has $m \downarrow 0$ and $\|D_A X^*\| \to \infty$ while $R_* \to I$.*

The proof is in Section A.13.

The construction separates the mean of two tokens from their contrast. Let $E = \mathbf{1}\mathbf{1}^\top/2$ project onto their common value and $Q = I - E$ onto their difference. For anchors $A = a(1, -1)^\top$, $0 < a < 1$, choose $0 < \mu < 1$ and set

$$k = \frac{\beta - m}{\alpha}, \qquad z_* = \frac{\beta a}{1 - \alpha\mu}, \qquad \mathcal{P}(X) = E + g\left(\frac{x_1 - x_2}{2}\right) Q, \tag{12.3}$$

where $g(z)$ is a smooth clipping of $\mu + k(z - z_*)$ to $[0, 1]$, equal to that affine function near z_*. Every entry of $\mathcal{P}$ is $(1 + g)/2$ or $(1 - g)/2$, so the builder is nonnegative and stochastic, with $L_P = k$. The unit ball is a common contraction domain, including a neighborhood of A. At the equilibrium $X^* = z_*(1, -1)^\top$,

$$R_* = E + \frac{\beta}{1 - \alpha\mu} Q, \qquad H = E + \underbrace{\frac{\beta}{1 - \alpha\mu - \alpha k z_*}}_{G} Q. \tag{12.4}$$

The weight derivative contributes the extra term $\alpha k z_*$ in the response denominator. As $a, \mu \uparrow 1$, the frozen contrast coefficient tends to one while $G \to \beta/m$. Each off-diagonal entry of H is $(1 - G)/2$, which becomes negative. Changing one anchor can redistribute positive weights enough to outweigh its direct contribution; Worked example 12.3 gives a numerical example on a certified common domain. Section A.13 specifies the smoothing and proves both limits.

Thus the bound (7.1) is optimal even at a fixed positive margin. For example, $\alpha = 1/2$ and $m = 1/4$ permit response approaching two with contraction factor at most $3/4$ and frozen smoother approaching identity. Unbounded response requires a vanishing margin. In the explicit diverging family of Proposition 12.2, at $\varepsilon = 0.01$, frozen contrast retention is 0.999900, response gain is 98.049, and the cross-token derivative is -48.524. Their linear-response neighborhood shrinks as the boundary is approached (Section A.13).

A worked example shows the mechanism in numbers (Bruce, 2026h).

Worked example 12.3 (Raising one input lowers another output)**.** Consider two tokens, each represented by one real number. A *token* is an indexed item; its *anchor* a_i is its original input, its *state* $x_i^{(t)}$ changes during iteration, and its *output* x_i^* is the converged state. In the image experiments of Section 12.4 a token is an image, with a vector of features instead of a single number. Here there is no classifier: the output is the number itself. Both tokens use the update

$$x_i^{(t+1)} = \underbrace{\tfrac{1}{2}a_i}_{\text{restart, coefficient } 1/2} + \underbrace{\tfrac{1}{2}\left[p_{i1}(X^{(t)})x_1^{(t)} + p_{i2}(X^{(t)})x_2^{(t)}\right]}_{\text{positive averaging of current states}}. \tag{12.5}$$

Each row of P sums to one. The coefficient on the token's own anchor is always $1/2$; the rule that builds P is also fixed. Its weights change because the states change. This is the diverging family of Proposition 12.2 (Section A.13) at $\varepsilon = 0.1$.

1. Raise only the first anchor. The original anchors $(a_1, a_2) = (1, -1)$ have equilibrium $(x_1^*, x_2^*) = (0.990099\ldots, -0.990099\ldots)$. Replace a_1 by 1.001, keep $a_2 = -1$, and hold these new anchors fixed throughout the second run. To expose the response, start from the old equilibrium. Contraction makes the final output independent of this initialization.

2. Let the same update rebuild its weights. The first update raises x_1 by 0.0005 and leaves x_2 unchanged. Their separation has increased. In this builder, larger separation reduces the weight each token assigns to the other's current state. From the next update onward, the second token receives less of the higher state. Repeated rebuilding continues this shift.

3. Compare the converged outputs. The second output falls to $-0.993947\ldots$: a decrease of $0.003848\ldots$, about 3.85 times the 0.001 increase in the first anchor. Holding the old equilibrium weights fixed instead gives a small *increase* of $0.0000049505\ldots$. The restart coefficient remains $1/2$ in both computations (Figure 12.1).

The weights. Let $z = (x_1 - x_2)/2$. The builder is

$$P(X) = \frac{1}{2}\begin{pmatrix} 1+g(z) & 1-g(z) \\ 1-g(z) & 1+g(z) \end{pmatrix}, \qquad g(z) = c_w\left(0.99 + 0.9\left(z - \frac{100}{101}\right)\right), \tag{12.6}$$

where $w = 0.0025$ and $c_w(u) = (2w)^{-1}\int_{-w}^{w} \mathrm{clip}_{[0,1]}(u+t)\,dt$. This continuously differentiable clipping keeps g in $[0,1]$ and equals its argument on $[w, 1-w]$. Both trajectories in Figure 12.1 stay in that affine region, with strictly positive weights. The builder depends on current states, not directly on the anchors.

One-step weights and effective anchor coefficients differ. The entry p_{21} weights the first token's *current state* in the second token's averaging term. At an equilibrium, the frozen smoother $R_* = \frac{1}{2}(I - \frac{1}{2}P_*)^{-1}$, with entries r_{ij}, instead records the total contributions of the *original anchors*, including all indirect paths. Thus

$$x_2^* = r_{21}a_1 + (1 - r_{21})a_2, \qquad r_{21} = \frac{1}{2}\left(1 - \frac{1}{2-g}\right).$$

For the adaptive runs, these quantities are:

Quantity	Original anchors	Raised first anchor
Direct restart coefficient on a_2	0.5	0.5
Equilibrium weight p_{21} on current state x_1	0.00500000	0.00304344
Effective coefficient r_{21} on original anchor a_1	0.00495050	0.00302503
Effective coefficient r_{22} on original anchor a_2	0.99504950	0.99697497
Converged second output x_2^*	−0.99009901	−0.99394692

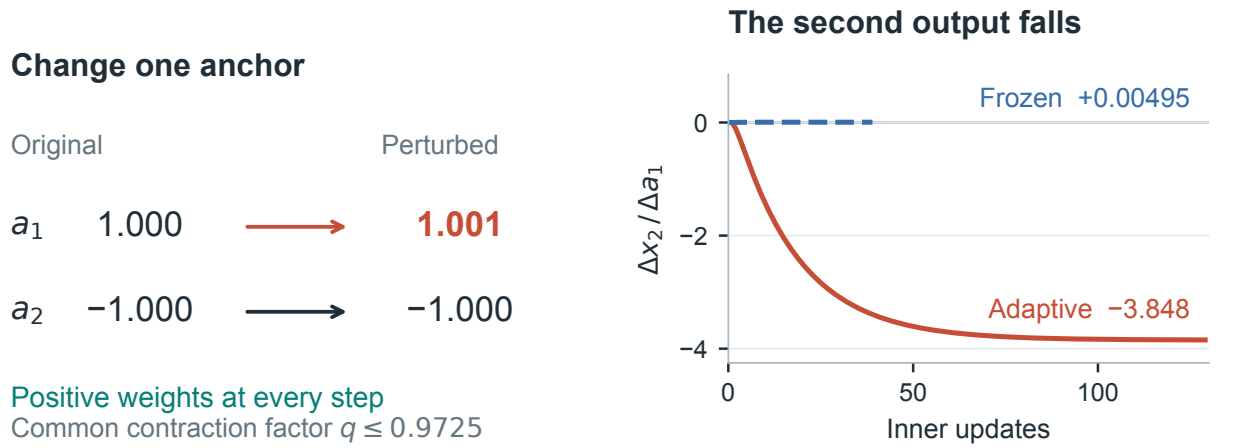


Figure 12.1: **A certified positive mixer has a negative cross-token response.** Raising only a_1 by 0.001 lowers the converged x_2 by 0.003848. Freezing the old equilibrium weights instead raises x_2 by 0.00000495. Both adaptive runs are contractive on the same domain. Curves start at the old equilibrium; the changed input persists at every update. This constructed witness supplies a global guarantee distinct from the locally checked denoising example in Figure 12.5.

Here *downweighting the first anchor* means reducing r_{21}. The second output relies more on its own, lower anchor through the changed averaging paths, while its direct restart coefficient stays fixed.

The negative change is a difference of positive averages. Write r_{21}^0 and r_{21}^1 for the two effective coefficients, and $\delta = 0.001$ for the increase in the first anchor. The output change decomposes exactly as

$$\Delta x_2^* = \underbrace{r_{21}^0 \delta}_{\text{raise the anchor at fixed weights}} + \underbrace{(r_{21}^1 - r_{21}^0)(1 + \delta - (-1))}_{\text{redistribute weight toward the lower anchor}}$$
$$\approx 0.0000049505 - 0.0038528608 = -0.0038479103.$$

The redistribution outweighs the direct positive contribution. Neither averaging matrix has a negative entry, and every state remains between its run's anchors. The negative quantity is the *response*: the change in the second output divided by the increase in the first anchor. The local derivative at the original anchors is $\partial x_2^*/\partial a_1 = -3.704829\ldots$; the finite change above gives $-3.847910\ldots$ because the map is nonlinear.

Why this is compatible with convergence. The builder has $L_P = 0.9$. Both anchor configurations lie in the common invariant ball $\max_i |x_i| \le 1.05$, where the contraction upper factor is $\frac{1}{2}(1 + 0.9 \times 1.05) = 0.9725 < 1$. Each run therefore has a unique attracting equilibrium. Contraction compares different state trajectories with the *same* anchors; the response compares equilibria with *different* anchors. A transient state perturbation with the anchors held fixed dies away. The changed anchor here persists at every restart.

The executable example of Bruce (2026h) checks the iteration against an 80-digit quadratic solution, the frozen comparison, the change decomposition and a central difference.

The sharp bound is compatible with almost unchanged frozen values (Figure 12.2) (Bruce, 2026h). At fixed contraction margins, smooth two-token families approach β/m. Near the convergence boundary, a separate family's adaptive response diverges while its frozen contrast retention approaches one. And across all 360 small-cloud confirmation configurations of dot-product (K0) and

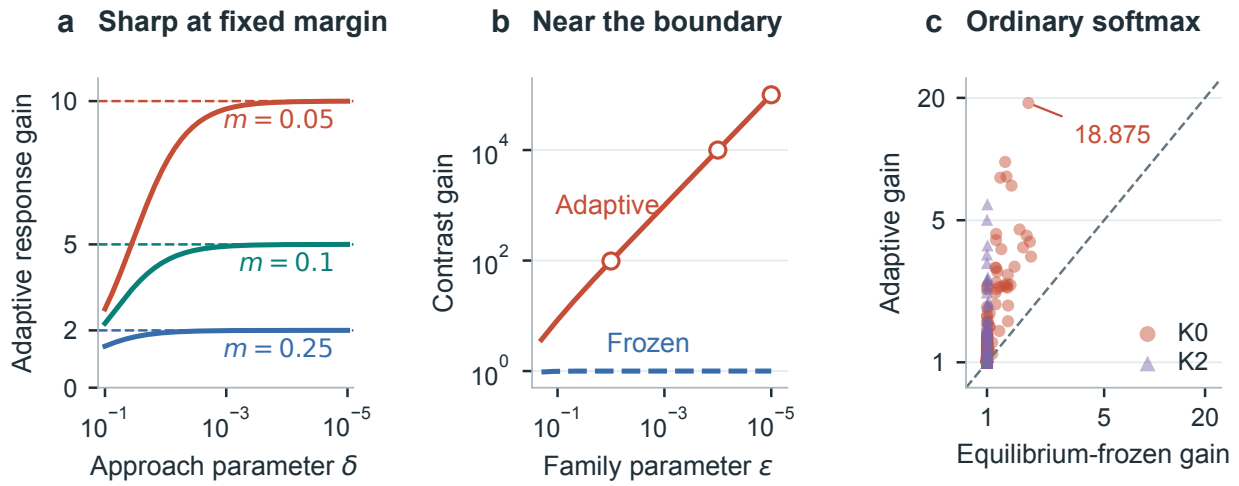


Figure 12.2: **The sharp response bound is compatible with almost unchanged frozen values.** (a) Smooth two-token families approach β/m at fixed contraction margins; dashed lines are the respective bounds, with $\alpha = \beta = 1/2$. (b) A separate family approaches the convergence boundary: adaptive response diverges while frozen contrast retention approaches one. Open circles are independent 80-digit finite differences. (c) All 360 small-cloud confirmation configurations for dot-product (K0) and adaptive-bandwidth Gaussian (K2) softmax (Section 12.2). The maximum Euclidean response is 18.875, versus 1.766 with its equilibrium weights frozen. The diagonal denotes equal gain. These designed configurations establish possibility, not prevalence.

adaptive-bandwidth Gaussian (K2) softmax, which Section 12.2 describes, the largest Euclidean response is 18.875, against 1.766 with its equilibrium weights frozen. These designed configurations establish possibility, not prevalence.

12.2 Ordinary softmax kernels settle while amplifying

Ordinary softmax also amplifies (Bruce, 2026h). Two builders are evaluated: tied dot-product weights (K0) and Gaussian weights whose bandwidth depends on the evolving cloud (K2). Both apply rowwise softmax; K2 uses logits

$$\ell_{ij}^{(2)} = -\frac{s\|x_i - x_j\|_2^2}{2h^2(X)}, \qquad h^2(X) = 0.01 + \frac{\sum_{i\neq j}\|x_i - x_j\|_2^2}{N(N-1)}. \tag{12.7}$$

A geometry study crosses four designed cloud families, isotropic, two-cluster, bridge and duplicate-pair, at $(N, d) = (16, 8)$ and $(64, 16)$, together with masked topology variants. Each of its two phases, discovery and confirmation, contains 900 configurations, and the phases use separate random draws. Every configuration is solved from the input and from zero; the 360 small-cloud configurations per phase also receive full-state derivative calculations. These grids test possibility and the scope of the certificate, rather than population frequency. The complete builder is differentiated, including its bandwidth. The geometry trajectories settle from both prescribed starts except for one confirmation solve from zero, whose residual is small but whose movement still exceeds the tolerance at its recorded iteration limit. Of the 900 confirmation configurations, 92 reach distinct endpoints. The

largest dense response is 18.875 despite $\rho(DF) = 0.9397$: a bridge cloud with K0, $N = 16$, $d = 8$, $\alpha = 0.6$, $s = 30$ in the confirmation grid. Freezing P_* gives 1.766; rebuilding once from each perturbed input and then holding the weights fixed gives 3.721. An independent central difference agrees with the adaptive derivative to relative error 3.97×10^{-6}.

These are Euclidean operator norms; a row-stochastic matrix can exceed one in that norm. Along the same selected direction, the maximum-row gain is 15.559 adaptive versus 0.821 frozen, so the separation also holds in the theorem's norm. The same check extends to all 360 small-cloud confirmation configurations: reconstruct H and evaluate each top Euclidean direction V in the maximum-row norm,

$$g_{\mathrm{row}}(V) = \frac{\|H[V]\|_{\infty,2}}{\|V\|_{\infty,2}} \le \|H\|_{\infty,2\to\infty,2} \le \max_i \sum_j \|H_{ij}\|_2,$$

where H_{ij} are $d \times d$ derivative blocks. The last inequality follows from the triangle inequality and bounds the true induced norm; the measured directional gain is a lower bound. Of the 360 designed configurations, 196 attain a gain above $1 + 10^{-8}$ in the tested direction. For the response maximum, the block-sum upper bound is 20.936. Every frozen smoother has induced maximum-row norm exactly one. The qualitative separation therefore survives in the norm of the theorem, without labeling a Euclidean singular vector as a maximizer of that different norm (Figure 12.3b). The example is an exploratory maximum. The full response inverse, not the scalar proxy $\beta/(1 - \rho(DF)) = 6.635$, determines its sensitivity.

The amplification also survives finite perturbation sizes. For the same response maximum, let V be the unit Frobenius-norm direction of greatest local response, and solve independently at $A \pm hV$ for thirteen log-spaced amplitudes $h \in [10^{-4}, 10^{-1}]$. The central gain $\|X^*(A+hV) - X^*(A-hV)\|_F/(2h)$ is 18.882 at $h = 10^{-4}$, close to the derivative's 18.875, with relative vector prediction error 3.98×10^{-4}. It is 19.690 at $h = 10^{-3}$, reaches 45.946 near $h = 1.78 \times 10^{-3}$ and remains 3.930 at $h = 0.1$, versus a frozen directional gain of 0.869 throughout. One-sided gains differ, reaching 77.373 on the negative side. Starting each perturbed solve from either its anchor or the nominal equilibrium gives gains agreeing within 2×10^{-9}, and every new-state residual is below 10^{-14}. Thus the amplification survives finite changes, although its size is nonlinear and the local derivative does not describe the entire excursion (Figure 12.3a). These calculations establish finite-amplitude amplification, not monotone decay or a uniform linear-response neighborhood, and they do not identify the nonlinear excursion as a particular bifurcation. Nor is the example hull-certified: the nominal anchor-hull contraction-margin bound is -31.058. A sampled trajectory norm or spectral gap cannot supply a common-neighborhood certificate or justify substituting a "local margin" into β/m.

Ordinary kernels can also settle at different endpoints (Figure 12.4). In 107 of the 900 discovery configurations and 92 of the 900 confirmation configurations, the input and zero starts both settle but differ by more than 10^{-4}. K0 contributes 69 in each phase; K2 contributes 38 and 23. No distinct-limit outcome is globally certified: none occurs where the radius-1.25 contraction certificate holds, and failure of the certificate alone does not predict a multiple-limit outcome. These are designed configurations, not independent samples of a population of attention layers. The largest dense discovery response is 9.823740 on a bridge/K2 cloud at $\alpha = 0.6$, $s = 30$; its equilibrium-frozen response is approximately 1.0035. Confirmation's 18.875 maximum is bridge/K0 at the same parameters. Both maxima were selected from their respective grids. Finite-difference checks validate the local calculations at these examples. Prespecified image directions attenuate in both phases: a large singular-direction gain in a designed cloud does not imply typical amplification in the image experiment.

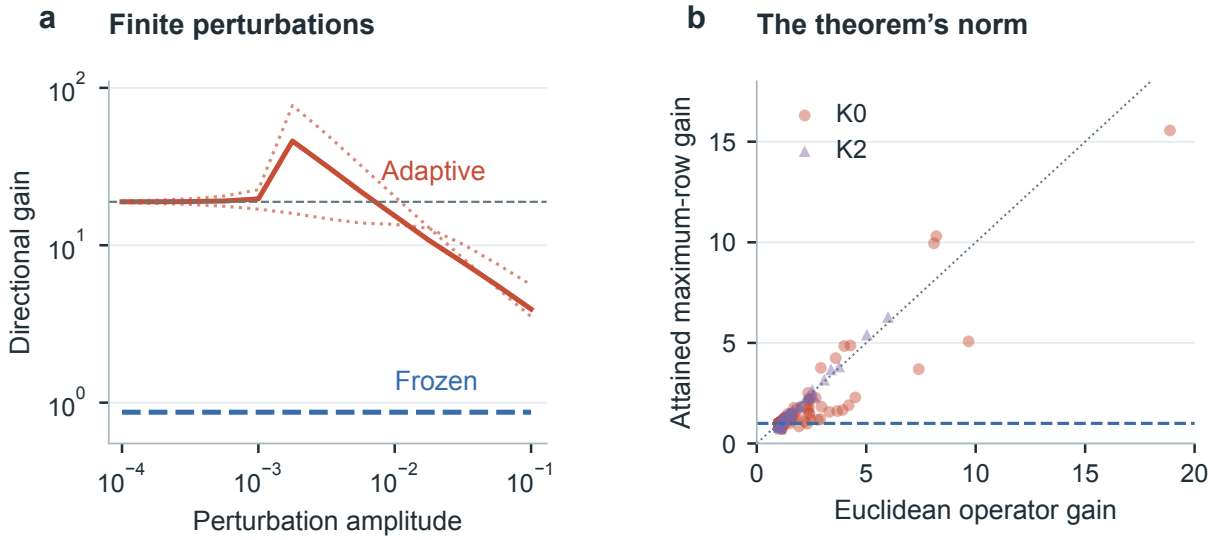


Figure 12.3: **Amplification persists beyond an infinitesimal Euclidean perturbation.** (a) Independently re-solved endpoints for the selected bridge/K0 maximum. Solid: central secant gain; dotted: one-sided gains. Frozen gain uses the same direction. (b) All 360 small-cloud confirmation cells, including identity controls. The vertical coordinate is an attained maximum-row gain along a Euclidean maximizing direction, not the induced maximum-row norm. Every frozen smoother has maximum-row operator norm one (dashed line). Both panels are exploratory diagnostics.

12.3 A photographic denoiser amplifies

A converged denoising loop can amplify its input (Figure 12.5) (Bruce, 2026h). In the photographic RED/NLM instance of Section 4.2, a non-local-means filter (4.1) rebuilds image-patch similarities from the evolving estimate while each fixed-point update retains the observed image (Romano et al., 2017; Reehorst & Schniter, 2019), with $\alpha = 0.95$ and bandwidth $h = 0.025$. The same photographic input gives distinct attracting limits from input and mean starts. A perturbation of at most 10^{-5} per pixel produces 19.12-fold amplification, versus 0.999 with equilibrium weights frozen. Its direction maximizes the local maximum-row gain (19.23); it is a selected probe, not sampled noise.

One observed image has two attracting states. Starting at A and at the constant image with value $\bar{A}$ gives endpoints separated by 0.241676 in maximum-pixel norm. Their update Jacobians have spectral radii 0.992583 and 0.990040, respectively, numerically establishing local attraction in the full 1,024-dimensional state space. Both endpoint residuals are below 1.5×10^{-15}, and their 100-update movements are below 2×10^{-13}. These are local checks at the computed roots, not a common-domain contraction certificate. The global certificate need not hold for an iteration to settle, and two settled runs need not agree.

A settled positive average can amplify a changed observation. At the input-start equilibrium, the full input-response matrix H of (12.2) has maximum-row gain 19.228965. Select its row r with largest absolute row sum and set $v_j = \operatorname{sign}(H_{rj})$. Then $\|v\|_\infty = 1$ and $(Hv)_r = \|H\|_\infty$; other output coordinates can respond negatively. Change the fixed observation to $A + 10^{-5}v$ and start at the old

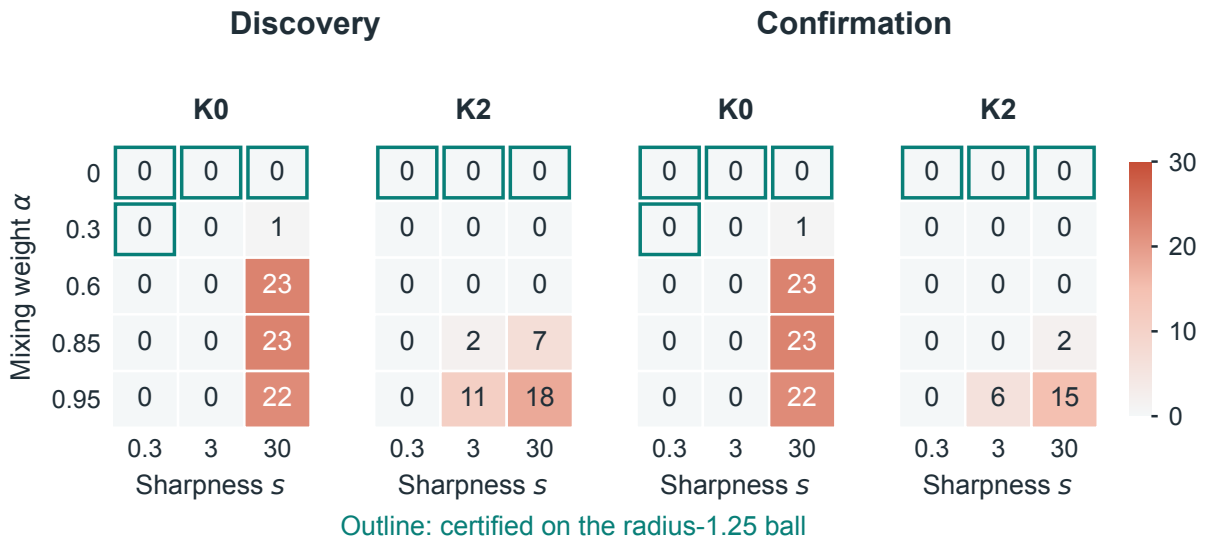


Figure 12.4: **Ordinary kernels can settle at different endpoints.** Counts of configurations where input and zero starts both settle but differ by more than 10^{-4}. Each cell contains 30 configurations. Outlined cells satisfy the radius-1.25 contraction certificate; none has distinct limits. Identity is included. The totals are 107/900 in discovery and 92/900 in confirmation. Failure of the certificate alone does not predict a multiple-limit outcome.

equilibrium. The converged maximum-pixel change divided by 10^{-5} is 19.119924. Holding the old equilibrium weights fixed gives 0.998500 in the same direction; its full operator gain is exactly one. The two calculations have the same initial state, changed anchor, and restart coefficient. Rebuilding accounts for the separation. The rebuilt response follows image structure, positive and negative, concentrated around image boundaries (Figure 12.6).

The direction is selected from the local derivative. It is not a random-noise sensitivity estimate. The crop was fixed before a six-setting full-resolution screen crossing $\alpha \in \{0.6, 0.8, 0.95\}$ and $h \in \{0.025, 0.05\}$; the focal case is selected for illustration. Table 12.1 reports that full screen and the preceding twelve-setting screen at half spatial resolution. The illustration neither estimates the frequency of amplification nor claims failure of standard tuned RED pipelines: the probe establishes possibility, and its checks at the computed roots are local, not a common contraction certificate. Its complete conventions, independent numerical checks and image provenance are given by Bruce (2026h).

12.4 Converged mixing can overturn class margins

Input sensitivity and classification ask different questions. Sensitivity compares nearby inputs; a class margin compares competing scores at one input (Bruce, 2026h). Cross-image aggregation can improve transductive classification, yet repeated feature re-representation reduces accuracy relative to one pass (Cui & Guo, 2021). Graph learning under heterophily likewise benefits from separating own and neighbor representations (Zhu et al., 2020). The role of adaptive rebuilding is isolated in test-time context mixing, the iteration of Section 4.2: fixed DINO ViT-S/16 features (Caron et al., 2021), fixed affine heads and CIFAR-10 (Krizhevsky, 2009), with images mixed within a context while the encoder and classifier stay unchanged. *A matched-query intervention.* From a

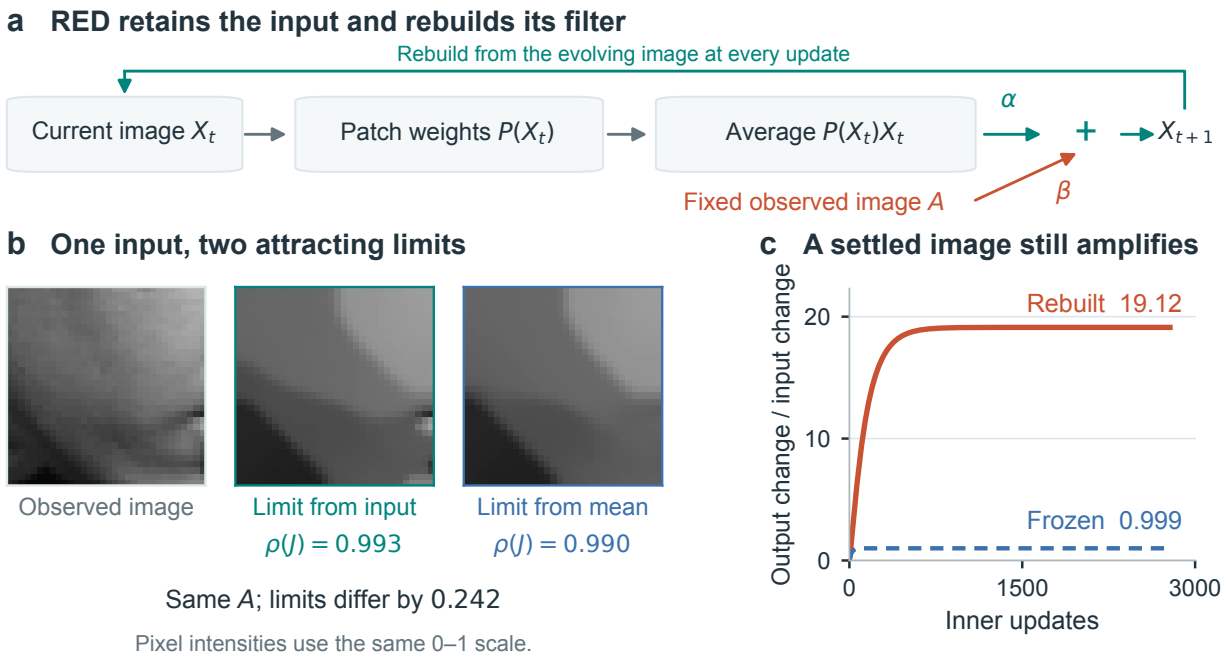


Figure 12.5: **A converged denoising loop can amplify its input.** (a) RED/NLM rebuilds patch similarities and retains the observed image (Romano et al., 2017). (b) The same photographic input gives distinct attracting limits from input and mean starts. (c) A perturbation of at most 10^{-5} per pixel produces 19.12-fold amplification, versus 0.999 with equilibrium weights frozen. Its direction maximizes the local maximum-row gain (19.23); it is a selected probe, not sampled noise. Here $\alpha = 0.95$, bandwidth 0.025. The figure's J is the update's Jacobian DF. Section 12.3 gives the derivative checks and the local scope, and Table 12.1 the full parameter screen.

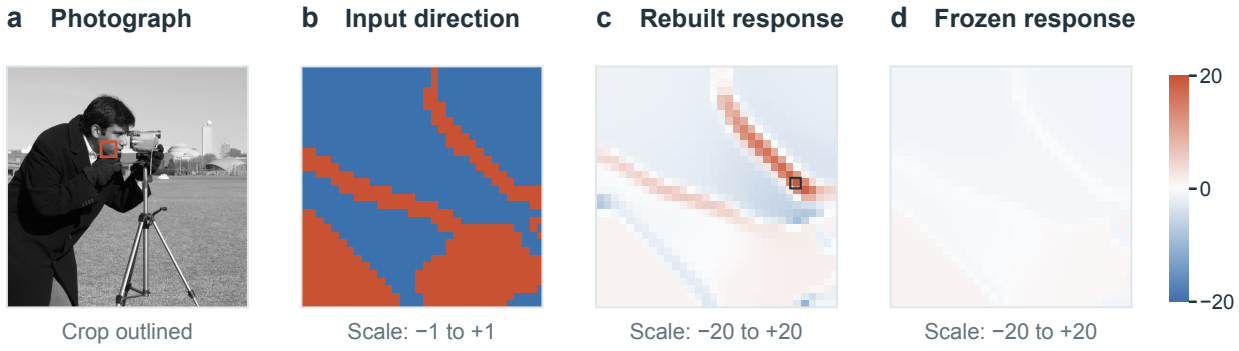


Figure 12.6: **The denoising response follows image structure.** (a) Source photograph with the fixed crop outlined (Lav Varshney, CC0). (b) The chosen input direction has entries ± 1, multiplied by 10^{-5} in the actual probe. (c–d) Equilibrium changes divided by 10^{-5}, using the same diverging color scale. Rebuilt weights produce both positive and negative responses, concentrated around image boundaries; the outlined pixel attains the largest absolute response. Frozen weights give gain below one. These maps are computed from the same runs as Figure 12.5; they do not depict additional image data or a random perturbation.

Table 12.1: **Complete RED/NLM exploratory screen.** Each setting uses the same crop at the indicated resolution and starts at the input A and its spatial mean $\bar{A}$. Both runs stop at maximum-pixel one-step movement 10^{-12} or an 8,000-step cap; every run below meets the movement test. Distances are maximum-pixel differences between the returned endpoints. The derivative quantities use the input-start endpoint, with DF the update's Jacobian; dashes indicate that they were not computed in the preliminary screen. The illustrated setting is 32×32, $\alpha = 0.95$, $h = 0.025$; the focal results in the text use tighter stopping criteria (Bruce, 2026h).

Resolution	α	h	Endpoint distance	Steps: $A/\bar{A}$	$\rho(DF)$	$\|H\|_\infty$
16×16	0.6	0.025	4.55×10^{-12}	64/76	—	—
16×16	0.6	0.05	6.56×10^{-12}	121/92	—	—
16×16	0.6	0.1	2.46×10^{-12}	47/52	—	—
16×16	0.6	0.2	1.87×10^{-12}	42/45	—	—
16×16	0.8	0.025	1.15×10^{-11}	191/195	—	—
16×16	0.8	0.05	7.82×10^{-13}	215/214	—	—
16×16	0.8	0.1	6.38×10^{-12}	96/99	—	—
16×16	0.8	0.2	6.32×10^{-12}	85/86	—	—
16×16	0.95	0.025	0.161596	1202/1413	—	—
16×16	0.95	0.05	0.118898	1458/376	—	—
16×16	0.95	0.1	3.70×10^{-11}	351/324	—	—
16×16	0.95	0.2	3.71×10^{-11}	333/325	—	—
32×32	0.6	0.025	7.84×10^{-13}	68/72	0.727023	2.440
32×32	0.6	0.05	7.22×10^{-12}	88/107	0.789770	3.386
32×32	0.8	0.025	0.069467	346/379	0.941305	9.479
32×32	0.8	0.05	4.62×10^{-12}	151/168	0.860332	3.129
32×32	0.95	0.025	0.241676	2707/1999	0.992583	19.229
32×32	0.95	0.05	0.065788	1375/2153	0.985207	11.879

documented extraction of all official images, full 384-dimensional features and a 16-dimensional PCA branch are prepared. Training-only centering, scaling, and the invertible map $z \mapsto z/\sqrt{1 + \|z\|_2^2}$ bound the states. Each branch fits its own ridge head. Each of 512 matched test queries receives 31 companions: three uniform draws or one deterministic nearest-neighbor context. Both branches retrieve in the training-fitted 16-dimensional features, without labels. Only queries are scored; all methods share companions. Uniform results average draws within query.

The focal parameters, $\alpha = 0.95$ and $s = 30$, were selected in the preceding image campaign and fixed before this study. Adaptive accuracy falls from 88.09% to 35.87% with uniform companions, a loss of 52.21 percentage points (pp; paired query-bootstrap 95% interval $[48.96, 55.53]$; accuracy differences are computed from exact counts before rounding). Building $P(A)$ once and freezing it retains 88.02%. These frozen weights are $P(A)$, rather than the weights of the adaptive equilibrium frozen in Sections 12.1 to 12.3; the latter would have the same nominal output as adaptive mixing and serve a different purpose in the response analysis. Each uniform draw exhibits the loss: adaptive accuracy ranges from 33.98% to 37.30%. Feature neighbors raise adaptive accuracy to 79.49%, but still leave an 8.59-point loss ($[5.27, 11.91]$); frozen weights retain 87.70%. Same-label fractions are 9.88% for uniform and 65.93% for nearest companions. Feature-based selection reduces interference without eliminating it. The intervals condition on this feature bank and context design (Figure 12.7).

No dimensional truncation is needed for this effect. The 16-dimensional branch also loses 46.48 pp. Table 12.2 places the full-dimensional result in the complete parameter comparison. At $\alpha = 0.85$,

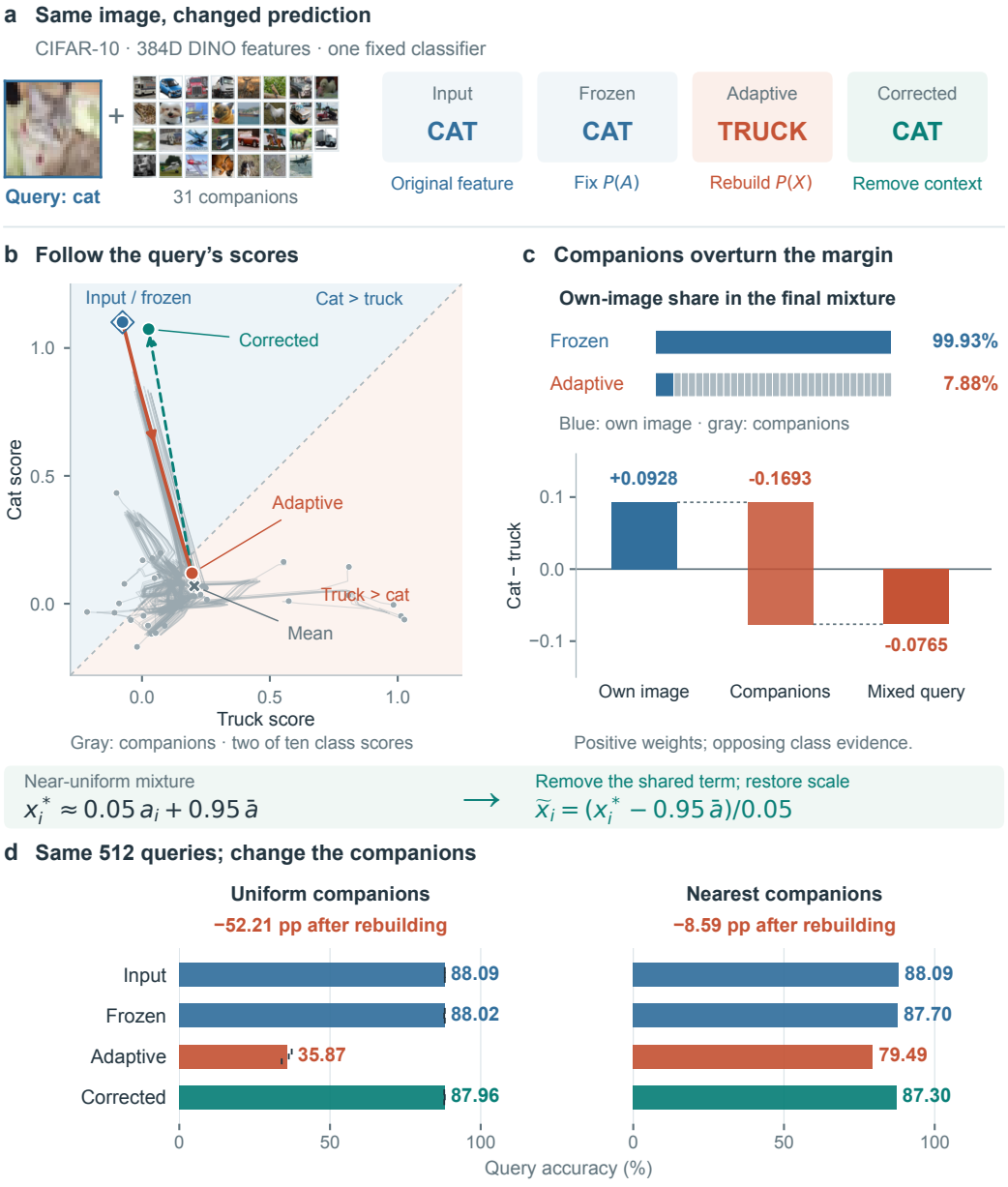


Figure 12.7: **Adaptive mixing hides class evidence that context removal recovers.** Full 384-dimensional DINO features, $\alpha = 0.95$, $s = 30$. (a) First stored query and 31 uniform companions; mixing acts on features. One ten-class head predicts cat with input or frozen weights, truck after rebuilding, and cat after correction. (b) Actual trajectories in two class scores; dashed arrow: correction, cross: original context mean. (c) Positive equilibrium-smoother weights carry opposing class evidence: companions outweigh the query's own cat-minus-truck contribution (12.10). (d) All 512 matched queries. Uniform accuracy averages three draws within query (marks show each draw); nearest companions use feature similarity without labels. Both losses are relative to input. The correction uses the mean $\bar{a}$ of all 32 original features (12.8): exact under uniform weights, diagnostic at the measured, approximately uniform equilibria. All focal solves settle.

Table 12.2: **The full-dimensional loss grows under stronger adaptive mixing.** Accuracy (%) on the same 512 queries; input accuracy is 88.09%. Uniform results average three draws within query. Frozen weights are built at the input. The first two settings are certified on the unit ball. The dagger marks one capped, unresolved solve retained in accuracy; all other adaptive solves settle. Correction at $(0.95, 30)$ reaches 87.96% with uniform and 87.30% with nearest companions.

		Uniform companions		Nearest companions	
α	s	Frozen	Adaptive	Frozen	Adaptive
0.30	0.3	88.28	88.28	87.89	87.89
0.60	0.3	87.37	87.30	84.96	84.96
0.85	30	88.09	76.11	88.48	80.27†
0.95	30	88.02	35.87	87.70	79.49

rebuilding already reduces uniform-context accuracy to 76.11%, while freezing retains 88.09%. Softer logits and stronger anchoring produce much smaller changes. The certified setting $(\alpha, s) = (0.6, 0.3)$, whose unit-ball global margin is $m = 0.4 - 2(0.6)(0.3) = 0.04$ (Section 7.1), loses 0.78 pp with uniform companions and 3.13 pp with neighbors; its adaptive and frozen accuracies are equal in three dimension and context combinations and differ by one query–context prediction in the fourth. Certification ensures convergence and uniqueness, not class-margin preservation. Both dimensions, all context draws and the paired intervals are given by Bruce (2026h).

The exact accounting of Section 12.5, which splits every output margin into a sample's own contribution and its companions', shows whose scores survive. In Figure 12.7, rebuilding reduces the query's own-image share from 99.93% to 7.88%; companion evidence overturns its cat-versus-truck margin. Across uniform contexts, mean direct self-weight falls from 99.85% to 3.41%; the equilibrium smoother retains 8.15% self-weight. Centered squared dispersion falls to 0.444% of its input value. Uniform weights $U = \mathbf{1}\mathbf{1}^{\top}/N$ have the exact smoother $\beta I + \alpha U$, which predicts self-weight $0.05 + 0.95/32 = 7.97\%$ and dispersion retention 0.25%; the modest discrepancy makes this an explanatory approximation, not an exact identification of the kernels. Of 1,536 query–context placements, 829 initially correct predictions become wrong and 27 initially wrong predictions become correct. Every loss has positive own-image evidence outweighed by companions (Figure 12.8). Counts refer to placements within the indicated design, not disjoint sets across repeated companion draws. This is exact accounting; the interventions test the explanation beyond the identity.

Removing the context term recovers the prediction. For uniform weights the exact smoother is $R_U = \beta I + \alpha U$ (Section A.13). Thus $x_i^* = \beta a_i + \alpha \bar{a}$, and centered squared dispersion retains $\beta^2 = 0.25\%$ at the focal setting. This suggests the label-free correction

$$\widetilde{X} = \frac{X^* - \alpha \mathbf{1}\bar{a}^{\top}}{\beta}, \qquad \bar{a} = \frac{1}{N}\sum_j a_j. \tag{12.8}$$

With the original head, corrected full-dimensional accuracy is 87.96%, a difference of -0.13 pp from the input ($[-0.78, 0.46]$); this is near recovery in the observed comparison, not a formal equivalence test. Nearest-context accuracy recovers to 87.30%. Together, the frozen-weight and context-removal controls identify accessible class evidence obscured by adaptive mixing. The correction uses the original context mean and amplifies errors by $1/\beta$; its present value is diagnostic, with no general reconstruction guarantee.

A refitted head does not substitute for the correction. Complementary readout controls use the preceding 16-dimensional feature bank, whose historical extraction provenance is incomplete. Across three complete 10,000-image test partitions, input accuracy is 75.33%, adaptive accuracy

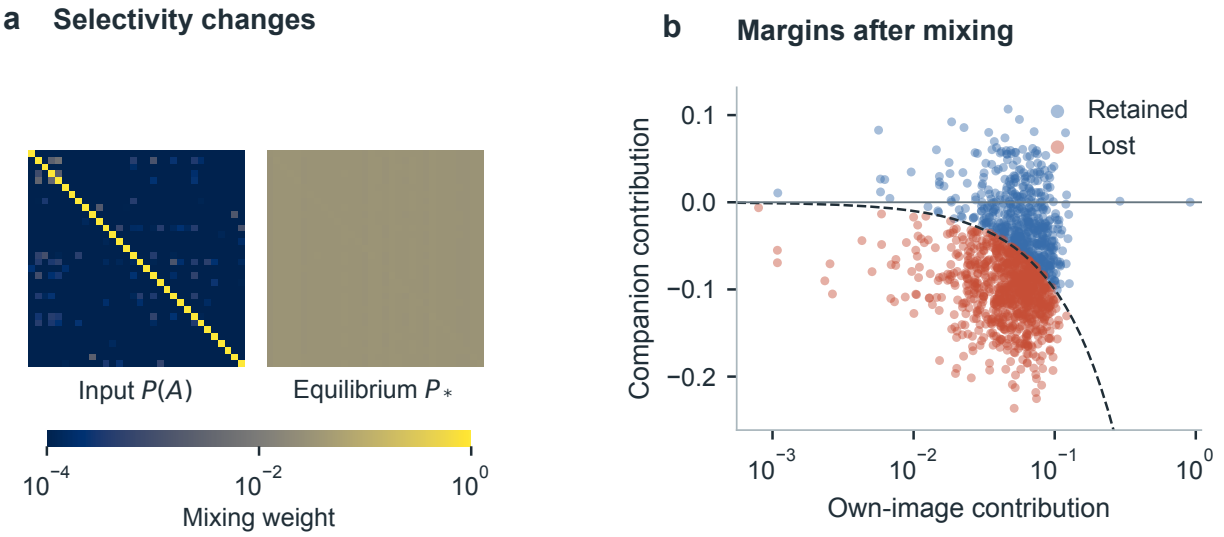


Figure 12.8: **Rebuilding changes whose class evidence reaches the output.** Full 384-dimensional DINO features with uniform companions, $\alpha = 0.95$, $s = 30$. (a) Initial and equilibrium kernels for the first query's context under the first draw, using the same logarithmic color scale: strong initial self-selectivity gives way to diffuse mixing. (b) Initially correct query–context placements across all three draws. The true class is compared with the strongest incorrect class at the adaptive endpoint; affine bias is allocated by smoother weights. Below the dashed boundary, negative companion contributions outweigh positive own-image contributions. The 829 lost placements are repeated-context outcomes among 512 queries. This is exact accounting; the interventions in Figure 12.7 test the mechanism.

29.41%, and accuracy after refitting an affine head on refined training examples 29.42%. Context correction reaches 73.93% with the original head (Bruce, 2026h). A shared per-image head cannot generally subtract an offset that varies with the companion set; the correction explicitly supplies that information.

Why initially selective weights can become diffuse. Evaluate the dot-product logits at $X_U = \beta A + \alpha \mathbf{1}\bar{a}^\top$. Up to a rowwise constant,

$$\ell_{ij}(X_U) = s\beta^2\langle a_i, a_j\rangle + s\alpha\beta\langle \bar{a}, a_j\rangle. \tag{12.9}$$

The pairwise coefficient falls from 30 to 0.075, while the second term favors the same keys for every query. Exploratory builder-only interventions support the role of weakened pairwise distinctions. In twelve confirmation contexts of the same historical 16-dimensional bank, centering the builder inputs barely changes accuracy (34.38% to 34.90%). Raising sharpness to $30/\beta^2 = 12{,}000$ restores the 76.56% input accuracy, with direct self-weight 99.74%; frozen weights recover equally well. This is recovery through nearly identity mixing. Equation (12.9) describes the near-uniform regime, not a universal basin or transition criterion.

All focal trajectories meet residual tolerance 10^{-13} and 100-step movement tolerance 10^{-11}. The full study retains one unresolved non-focal solve among 16,384; its capped prediction remains in the reported accuracy. Even without an adaptive contraction certificate, the residual $r = X - \alpha PX - \beta A$

of a candidate endpoint X, at its realized frozen weights P, bounds the value-reconstruction error by $\|X - RA\|_{\infty,2} \le \|r\|_{\infty,2}/\beta$, where $R = \beta(I - \alpha P)^{-1}$ (Section A.13). At the focal setting this is at most 2×10^{-12}.

12.5 Own and companion margins

Exact accounting of whose scores survive. For an affine classifier with weights W and bias b, let $Z_0 = AW + \mathbf{1}b^\top$ be the scores of the anchors. Since $R_*\mathbf{1} = \mathbf{1}$ (Proposition 12.1), a converged average passes the scores through the same smoother, and every class margin splits exactly (Bruce, 2026h).

Proposition 12.4 (Own and companion margins)**.** *At an equilibrium, the output scores satisfy $Z^* = R_*Z_0$. A true-class-versus-competitor margin is therefore*

$$Z^*_{iy} - Z^*_{ik} = \underbrace{(R_*)_{ii}(Z_{0,iy} - Z_{0,ik})}_{\textit{own-image contribution}} + \underbrace{\sum_{j \neq i}(R_*)_{ij}(Z_{0,jy} - Z_{0,jk})}_{\textit{companion contribution}}. \tag{12.10}$$

The proof is in Section A.13.

The identity also gives a positive criterion for retaining a decision.

Corollary 12.5 (A sufficient condition for keeping a decision)**.** *For an initially correct query i of class y, define $b_{ik} = \max_{j \neq i}[Z_{0,jk} - Z_{0,jy}]_+$, where $[u]_+ = \max(u, 0)$. Its class remains correct whenever, for every competitor k,*

$$(R_*)_{ii}(Z_{0,iy} - Z_{0,ik}) > (1 - (R_*)_{ii})b_{ik}. \tag{12.11}$$

The proof is in Section A.13. The right side bounds adverse companion evidence. The guarantee requires task margins and context information beyond stochasticity or contraction.

Chapter 0 states that the identity holds at every iteration of mixing initialized at the anchors, as well as at the fixed point (Section 0.4.3); for a mixing matrix R whose row mass $\rho_i = \sum_j R_{ij}$ is below one, it gains an affine-bias term, $(1 - \rho_i)(b_y - b_c)$, with b_y and b_c the readout's biases for the true and the rival class, as the construction of Figure 12.9 shows. The proof is in Section A.13; the full treatment is left for later work. On test-time mixing over real image contexts, a post hoc numerical audit verified the accuracies at all 23,040 endpoints (11,520 contexts) and reconstructed the saved scores of every settled endpoint, all but one, from the anchors' scores through the identity $Z^* = R_*Z_0$ to within 1.8×10^{-10}; in 48 precision replays no prediction changed.

Nonnegative context mixing can overturn a correct prediction, and a mixing matrix that loses row mass also requires an affine-bias correction (Figure 12.9). An exact illustrative construction holds four scalar anchors and a binary affine readout fixed: $a = (1/2, -3/2, -3/2, -3/2)$, $z_y(x) = x + 1/2$ and $z_c(x) = 0$. The focal sample's class margin is then $m_i = +1$ (a class margin, not the contraction margin m), each companion's is -1, and $d = b_y - b_c = 1/2$. The row $(3/4, 1/12, 1/12, 1/12)$ of the mixing matrix R preserves the correct decision: $3/4 - 1/4 = 1/2$. Uniform mixing gives $1/4 - 3/4 = -1/2$, overturning it. Halving that row gives $1/8 - 3/8 + (1 - 1/2)(1/2) = 0$: a tie, despite a negative sum of own and companion terms. These prescribed rows illustrate the exact affine identity, including its sub-stochastic extension; they are not measured endpoints or evidence of an abstention repair.

The degradation of Section 12.4 does not need a directed kernel (Bruce, 2026h).

Proposition 12.6 (Reversible kernels and an invertible smoother)**.** *For symmetric nonnegative affinities $K = K^\top$ with positive row sums d_i, the kernel $P_{ij} = K_{ij}/d_i$ has stationary distribution*

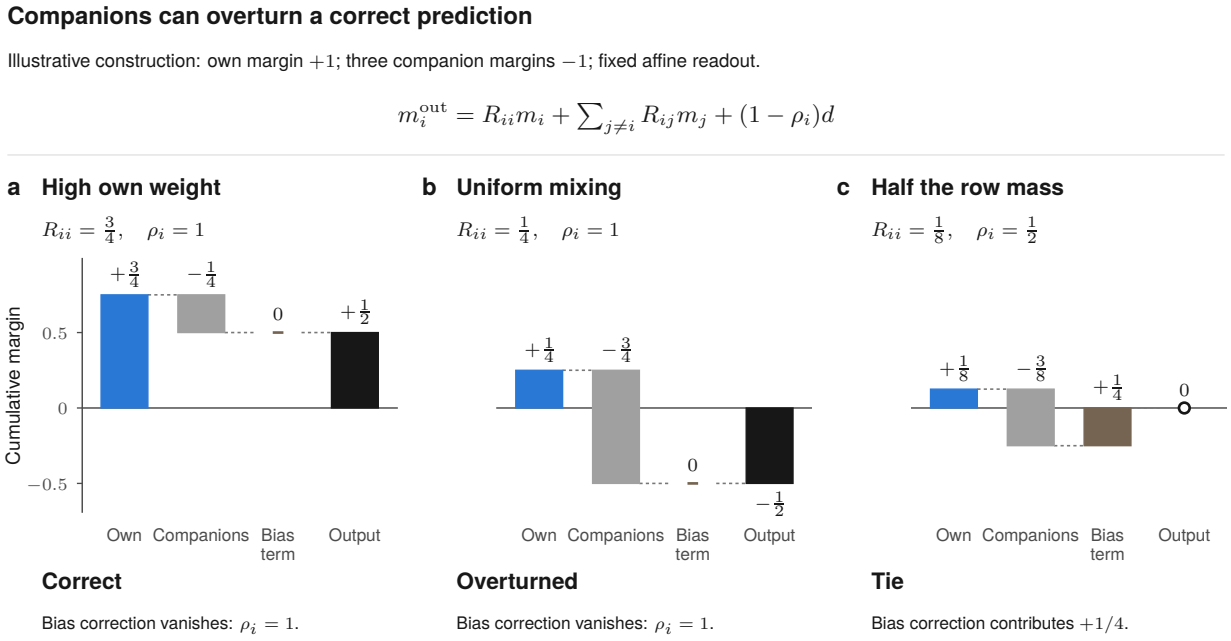


Figure 12.9: **Nonnegative context mixing can overturn a correct prediction; losing row mass also requires an affine-bias correction.** An exact illustrative construction holds four scalar anchors and the binary readout fixed: $a = (1/2, -3/2, -3/2, -3/2)$, $z_y(x) = x + 1/2$, and $z_c(x) = 0$. Thus the focal margin is $m_i = +1$, each companion margin is -1, and $d = b_y - b_c = 1/2$. Bar heights are signed contributions; dashed joins track their cumulative sum, on identical vertical scales. Black bars show the output margin, and the open circle denotes zero. (a) The row $(3/4, 1/12, 1/12, 1/12)$ preserves the correct decision: $3/4 - 1/4 = 1/2$. (b) Uniform mixing gives $1/4 - 3/4 = -1/2$, overturning it. (c) Halving that row gives $1/8 - 3/8 + (1 - 1/2)(1/2) = 0$: a tie, despite a negative sum of own and companion terms. Here $\rho_i = \sum_j R_{ij}$ is the row mass of the mixing matrix R. These prescribed rows illustrate the exact affine identity, including its sub-stochastic extension; they are not measured endpoints or evidence of an abstention repair.

$\pi_i = d_i / \sum_j d_j$ *and satisfies detailed balance,* $\pi_i P_{ij} = \pi_j P_{ji}$. *For fixed* P, *the smoother* $\beta(I - \alpha P)^{-1}$ *is invertible.*

The proof is in Section A.13. Unmasked K0 and K2 have this property at every state, since their logits are symmetric in i and j; they are reversible in the sense of Section 1.2. The observed image degradation therefore coexists with detailed balance, and margin reversal under a fixed readout is compatible with recoverable representations.

Across the confirmation image grid, strong mixing lowers mean accuracy, and adaptive rebuilding adds a smaller loss relative to input-frozen mixing (Figure 12.10). The grid compares the dot-product softmax kernel K0 and the squared-distance softmax kernel K2 with state-dependent bandwidth at five measured mixing weights, $\alpha \in \{0, 0.3, 0.6, 0.85, 0.95\}$. Each builder and mixing weight includes 576 contexts of 32 images: CIFAR-10 and CIFAR-100; DINO ViT-S/16 and SimCLR ResNet-18 features (Caron et al., 2021; Chen et al., 2020b); 16-dimensional feature states and belief states, the latter class-probability vectors; uniform and feature-neighbour contexts; logit scales $s \in \{0.3, 3, 30\}$; and four episodes per seed and setting, over three confirmation seeds. Lines are equal-weight grid means, and bands span the three seed means, an observed range rather than a confidence interval;

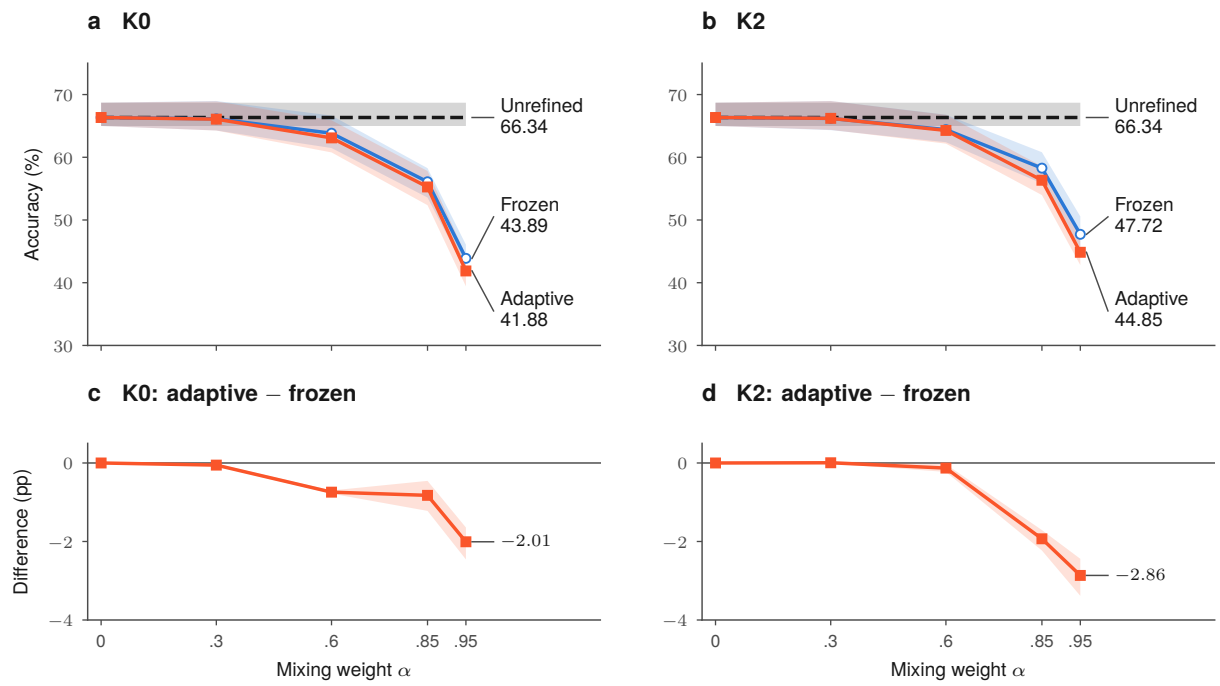


Figure 12.10: **Strong mixing lowers mean accuracy across the confirmation image grid, and adaptive rebuilding adds a smaller loss relative to input-frozen mixing.** (a,b) Unrefined, input-frozen and adaptive accuracy for the dot-product softmax kernel K0 and the squared-distance softmax kernel K2 with state-dependent bandwidth; frozen uses open circles, adaptive filled squares and unrefined a dashed reference. (c,d) Paired adaptive-minus-frozen accuracy differences in percentage points (pp) at the same settings. Lines are equal-weight grid means; bands span the three confirmation seed means (seeds 100–102), an observed range rather than a confidence interval. Each builder and mixing weight includes 576 contexts of 32 images: CIFAR-10/100, DINO ViT-S/16 and SimCLR ResNet-18, 16-dimensional feature states and belief states, uniform and feature-neighbour contexts, logit scale $s \in \{0.3, 3, 30\}$, and four episodes per seed and setting. Contexts and repeated image placements are dependent. The five measured mixing weights are $\alpha \in \{0, 0.3, 0.6, 0.85, 0.95\}$; connecting lines are guides. This descriptive fixed-grid comparison does not select policies or test the pending repairs.

contexts and repeated image placements are dependent. The grid is a historical one whose checkpoint and preprocessing provenance is incomplete (Bruce, 2026h). This descriptive fixed-grid comparison selects no policy and tests no repair.

Pending: Independently selected adaptive and frozen policies, compared on the full image test set, on graphs over confirmation seeds and on memory streams.

Abstention meets this identity as well. A head that can abstain makes the mixing sub-stochastic, and the questions it raises there, the loop gain of a sub-stochastic softmax, the floor on a sample's

own share at collapse and the failure of a fixed null across temperatures, make up the prediction that Section 8.6 records; they are not repeated here.

12.6 Anatomy of a focal failure

The accounting can be followed context by context and placement by placement (Figure 12.11). In a post hoc focal setting, CIFAR-10 with 16-dimensional prepared DINO ViT-S/16 features, uniform 32-image contexts , the dot-product kernel K0, $\alpha = 0.95$ and $s = 30$, discovery and confirmation compare input-frozen and adaptive mixing on the same contexts: frozen holds $P(A)$ fixed, and adaptive rebuilds $P(X)$. Seeds 0–2 and 100–102 each supply four contexts, so each phase has 12 contexts and 384 placements; unrefined accuracy is 73.44% in discovery and 76.56% in confirmation. For every context the figure gives the kernel and resolvent self-weights, the mean diagonals of P and of $R = (1 - \alpha)(I - \alpha P)^{-1}$; the other-label weight, the mass of R on differently labeled members; the retained dispersion; and the accuracy. Adaptive rebuilding collapses self-weight and lowers accuracy, and its resolvent self-weight can be read against the own share that uniform mixing leaves at collapse, $\beta + \alpha/N = 0.05 + 0.95/32 = 7.97\%$ (Section 12.4). The contexts come from the same historical image grid as Figure 12.10, whose provenance is incomplete, and the confirmation contexts are those of the builder probes of Section 12.4. A placement is a sample occurrence in a context, not an independent image; the displayed spread is descriptive context variation, not a confidence interval; and repair tests remain pending.

Every overturned initially correct placement keeps a positive own contribution that a larger negative companion contribution defeats (Figure 12.12). Initially correct means that the unrefined affine head predicts the true class. Of the 282 and 294 initially correct placements in discovery and confirmation, input-frozen mixing overturns 24 and 19, and adaptive rebuilding 200 and 172. The two contributions are the terms of (12.10), with the rival class taken as each method's highest-scoring non-true class at its endpoint; rivals can differ between methods, so paired placements need not represent the same class contrast. Here R is row-stochastic, so the affine-bias correction vanishes. This is descriptive focal accounting, with no interval estimated; repair tests remain pending.

12.7 When mixing helps: guidance

Pending: When mixing helps: the gain from propagation and the rate at which companions overturn correct predictions, against label agreement set at random, under frozen and under adaptive mixing.
Pending: Repairs for the focal failure, each fixed in advance: accuracy and resolvent self-weight against mixing strength for each repair, in the focal setting and in ordinary settings.

Chapter 0 states the guidance that follows for practice (Section 0.6.1): read test-time context mixing with an affine readout through each prediction's own and companion margins (Section 12.5). Its full treatment, as guidance for context-mixing methods, is left for later work.

12.8 Beyond cached features: prevalence

Pending: Patches, text and trained attention: signed class-margin influence maps and boundary F1 under frozen, adaptive and repaired patch mixing.
Pending: How often companions overturn correct predictions in practice: flip accounting across context-mixing methods, with the net gain, the harmful and helpful flips, and the share of harmful flips that companions overturn.

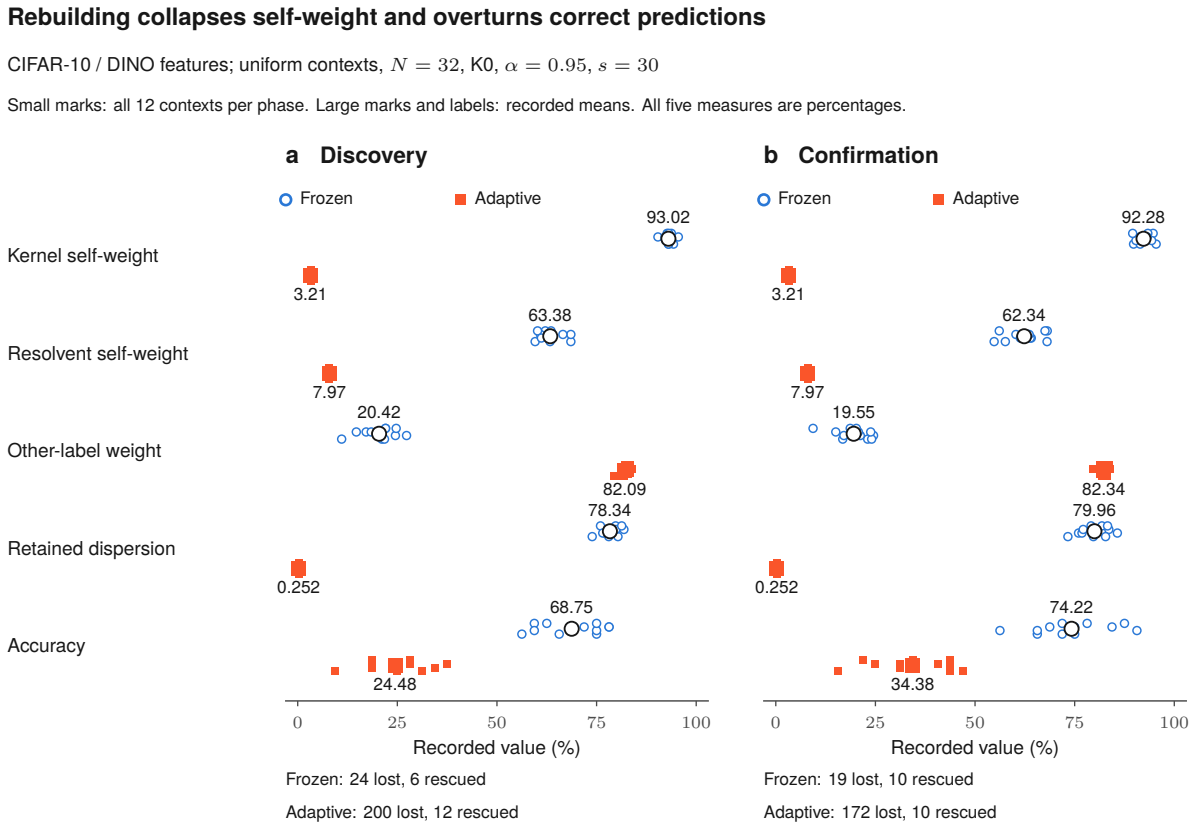


Figure 12.11: **Adaptive rebuilding collapses self-weight and lowers accuracy in this post hoc focal setting.** (a) Discovery and (b) confirmation compare input-frozen and adaptive K0 mixing on the same contexts. Frozen holds $P(A)$ fixed; adaptive rebuilds $P(X)$. Small open circles and filled squares show all 12 context values per method; large symbols and labels show the recorded phase means (overlapping marks are retained). Kernel and resolvent self-weights are mean diagonals of P and $R = (1-\alpha)(I-\alpha P)^{-1}$; at $P(X^*)$, R is the frozen smoother R_*. Other-label weight is the mass of R on differently labeled members. Retained dispersion is $\|X - \bar{X}\|_F^2 / \|A - \bar{A}\|_F^2$, expressed as a percentage. Unrefined accuracy is 73.44%/76.56% in discovery/confirmation. Lost and rescued counts are correct-to-wrong and wrong-to-correct placements. CIFAR-10 with 16-dimensional prepared DINO ViT-S/16 features, uniform 32-image contexts, $\alpha = 0.95$, $s = 30$; seeds 0–2 and 100–102, four contexts per seed and 384 placements per phase. A placement is a sample occurrence in a context, not an independent image. The displayed spread is descriptive context variation, not a confidence interval. Repair tests remain pending.

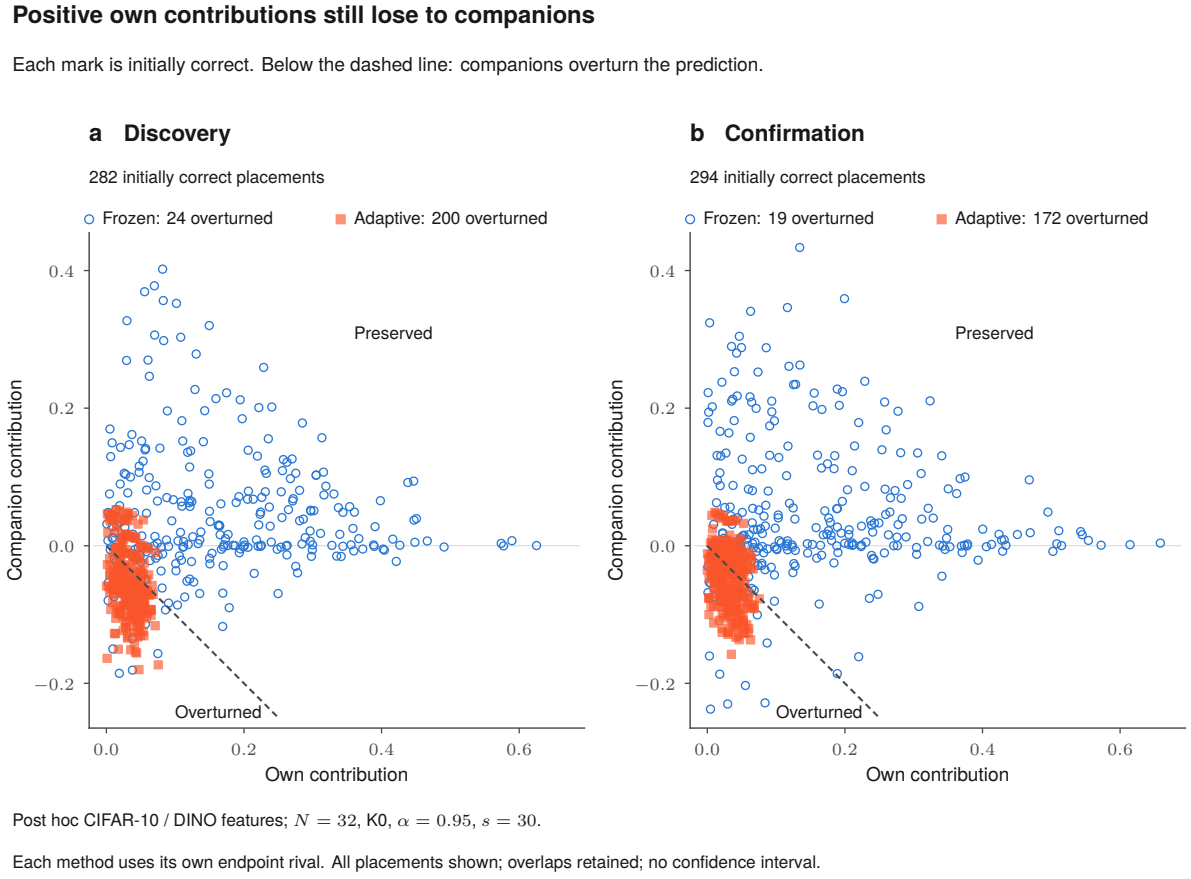


Figure 12.12: **Every overturned initially correct placement keeps a positive own contribution that a larger negative companion contribution defeats.** (a) Discovery and (b) confirmation show the same initially correct cohort under input-frozen mixing (open circles) and adaptive rebuilding (filled squares). Initially correct means that the unrefined affine head predicts the true class. For each point, c is that method's highest-scoring non-true class at its endpoint: the axes give $R_{ii}(z^0_{iy} - z^0_{ic})$ and $\sum_{j \neq i} R_{ij}(z^0_{jy} - z^0_{jc})$, where z^0 denotes initial scores and y the focal true class. Their sum is the output margin; the dashed line is zero. Here $R = (1-\alpha)(I - \alpha P)^{-1}$ is row-stochastic, so the affine-bias correction vanishes. Frozen uses $P(A)$; adaptive uses $P(X^*)$. Rivals can differ between methods, so paired placements need not represent the same class contrast. Post hoc focal setting: CIFAR-10 with 16-dimensional prepared DINO ViT-S/16 features, uniform contexts of 32, K0, $\alpha = 0.95$, $s = 30$. Discovery seeds 0–2 and confirmation seeds 100–102 supply four contexts per seed and 384 placements per phase; 282/294 are initially correct. Frozen overturns 24/19, adaptive 200/172. A placement is a sample occurrence within a context, not an independent image. All qualifying observations are shown with overlap retained; no confidence interval is estimated. This is descriptive focal accounting; repair tests remain pending.

12.9 What holds in trained networks

Exact for every network of the stated class. Own and companion margins (Section 12.5). An identity for any mixing followed by an affine readout, at every iteration and at the fixed point.

Measured in trained networks. Converged averages that amplify, and converged mixing that overturns class margins (Sections 12.1 to 12.4). Measured in iterated systems that apply one builder repeatedly (softmax kernels, a real denoiser, test-time mixing over 32-image contexts); the maxima were selected from grids.

What a converged average guarantees, matched to the quantity that is used, and the scope of these results are given once, in Table 7.2 and the paragraph after it (Section 7.8).

Chapter 13

Settled is not solved

Chapter 12 found settled averages that amplify their inputs and that settle at different endpoints from different starts (Sections 12.1 to 12.3). Both are cases of a more general point, which Chapter 0 states in brief (Section 0.4.3): a computation that settles to machine precision certifies neither its response, nor its uniqueness, nor that it attracts, nor that a loop built on it is stable. This chapter separates the four questions. Section 13.1 sets out the four readings of one fixed-point computation. Sections 13.2 to 13.4 show a settled solve that is under-resolved, settled endpoints that depend on the start, and a settled root that does not attract. Sections 13.5 and 13.6 turn to an outer loop that uses a finite computation, Section 13.7 to what an exponential moving average changes, and Section 13.8 to two loops that share state. Section 13.9 holds the measurements still to come in trained systems, Section 13.10 gathers the checks, and Section 13.11 says what of this holds in trained networks. The results are proved and shown on constructions and small systems; trained looped and implicit models are still to be tested.

13.1 Four readings of a fixed-point computation

One fixed-point computation can be read four ways: by its response to the input, by the uniqueness of its endpoint, by whether that endpoint attracts, and by the stability of an outer loop that uses it. Chapter 0 states that settling to machine precision certifies none of the four, so that each reading needs its own check (Section 0.4.3); their full treatment is left for later work.

For anchored adaptive averaging, Table 7.2, printed with the closing of Chapter 7, matches four guarantees to the objects that answer them, from the margin m for convergence to the finite derivative H_K for downstream feedback (Bruce, 2026h). Section 13.5 develops the finite derivative.

13.2 Settled but under-resolved

A float64 solve can settle and still be under-resolved (Bruce, 2026h). In the explicit diverging family of Section 12.1 at $\varepsilon = 10^{-4}$, a float64 run can satisfy residual 4.97×10^{-13} while its endpoint error remains 9.94×10^{-9}. A central difference with $h = 10^{-13}$ then reports gain about 55.83 instead of 9,998.00. Independent 80-digit roots recover the exact derivative, with relative errors between 2×10^{-14} and 2×10^{-10} across the three amplitudes; the independent finite-difference checks in Section 12.1 use these validated roots, not unresolved small differences. All 24 historical focal image contexts of Section 12.4 are also rerun from the input and the saved endpoint with independent float64 kernels and a 10^{-14} movement criterion, and all 48 replays retain every predicted label.

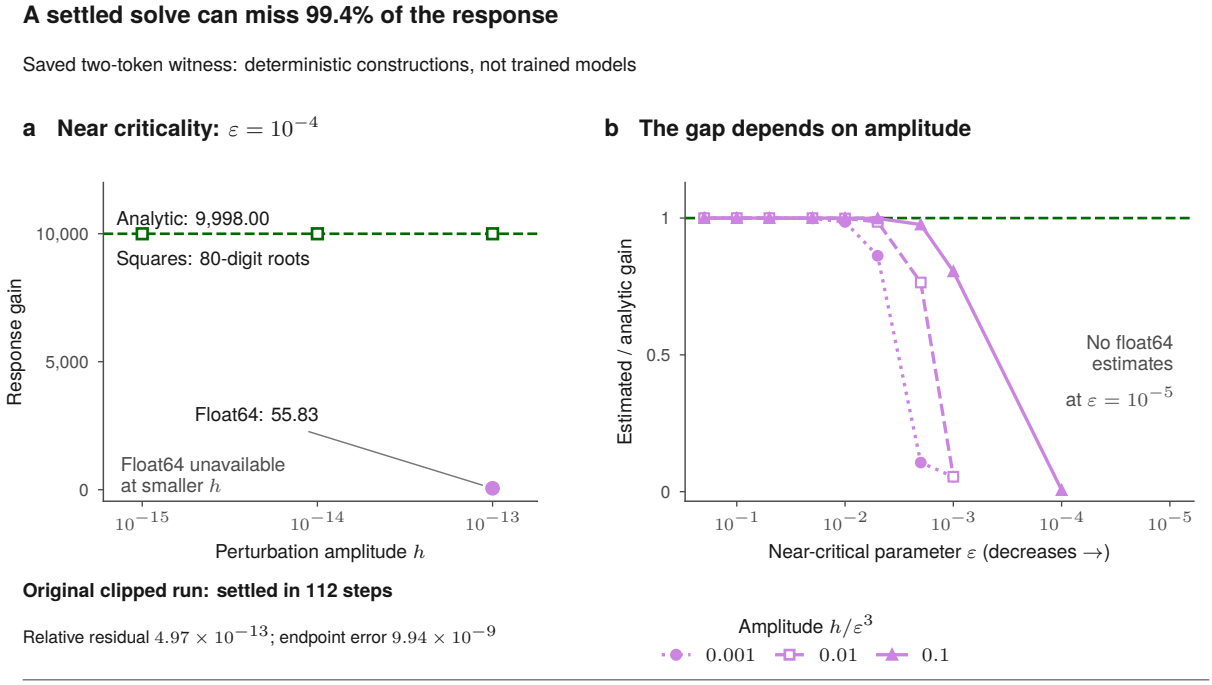


Figure 13.1: **A settled solve can miss 99.4% of the response.** Saved numerical evaluations of the two-token illustrative construction, the explicit diverging family of Section 12.1 (near-critical parameter ε, anchor perturbation amplitude h, response gain of the token contrast). **(a)** At $\varepsilon = 10^{-4}$, the float64 central difference at $h = 10^{-13}$ gives 55.83, against the analytic gain 9,998.00 (dashed green line); open squares are independently saved 80-digit root calculations at the three amplitudes. The residual and endpoint error below refer to the original unperturbed clipped run, not the perturbed solves. **(b)** Estimated gain divided by analytic gain across all ten scalar ε values, at $h/\varepsilon^3 \in \{0.001, 0.01, 0.1\}$. Clipped and smooth-clip gains coincide exactly, so each mark represents both; lines only join saved points. Float64 estimates are unavailable at all three amplitudes for $\varepsilon = 10^{-5}$ and the two smaller amplitudes for 10^{-4}. Every recorded perturbed solve is labelled settled; the original runs at $\varepsilon = 0.001$ and 0.002 reach the 4,096-step limit unresolved. There are 20 deterministic scalar cases (one per parameter and clipping form), with no sampling intervals; embedding replicates are outside this panel.

These checks address different numerical questions: resolving a tiny change for differentiation and preserving a classification decision under a tighter solve.

A settled solve can miss 99.4% of the response (Figure 13.1). The figure draws the saved numerical evaluations of the same illustrative two-token construction: the response gain of the token contrast, at near-critical parameter ε and anchor perturbation amplitude h. At $\varepsilon = 10^{-4}$ the float64 central difference at $h = 10^{-13}$ gives 55.83, against the analytic gain 9,998.00, and the residual and endpoint error quoted above refer to the original unperturbed clipped run, not to the perturbed solves. Across all ten scalar values of ε, at the amplitudes $h/\varepsilon^3 \in \{0.001, 0.01, 0.1\}$, the figure divides each float64 estimate by the analytic gain. Every recorded perturbed solve is labelled settled; yet float64 estimates are unavailable at all three amplitudes for $\varepsilon = 10^{-5}$ and at the two smaller amplitudes for 10^{-4}, and the original runs at $\varepsilon = 0.001$ and 0.002 reach the 4,096-step limit

unresolved. These are 20 deterministic scalar cases, one per parameter and clipping form, with no sampling intervals.

13.3 A settled state need not be unique

Settling from one start says nothing about where another start settles. Ordinary softmax kernels already show it: starts at the input and at zero can both settle and still end apart, none of them in a cell that satisfies the contraction certificate (Section 12.2); and the photographic denoiser reaches distinct limits from two starts, each attracting by local checks at its computed root (Section 12.3) (Bruce, 2026h).

Read by certificate class, the same designed survey shows distinct settled endpoints only outside the saved contraction certificate (Figure 13.2). The certificate is the sufficient bound $q = \alpha(1 + RL_P) < 1$ of Theorem 7.1 at the state-domain radius $R = 1.25$, with L_P the analytic Lipschitz upper bound for the builder; in each phase it holds in the identity cells, which do not mix, and in the nonidentity cells that Table 7.1 counts at that radius. A positive flag requires both starts, at the input anchor and at zero, to be labelled settled and their endpoints to differ by more than 10^{-4} in the maximum-row norm. The figure also breaks the flags down by builder and sharpness, where every flag of the dot-product builder falls at the sharpest logits, and by geometry. Each phase has 900 cells, over both builders, three sharpnesses and five mixing weights. The printed counts are descriptive summaries of a complete grid, not prevalence estimates or confidence intervals, and cells that share a source seed are not independent replicates.

13.4 A root is not an attractor

A trajectory can settle at a root that is not locally attracting (Figure 13.3). The illustrative construction is the linear update $F(x) = DF\,x$ on $\mathbb{R}^2$, with Jacobian $DF = \mathrm{diag}(1/2, 5/4)$ and root 0. From $(1, 0)$ the iterates are $x_t = (2^{-t}, 0)$, and the residual tends to zero. From $(1, \pm 1/32)$ the transverse displacement is $\pm(5/4)^t/32$ and grows at every step. The Jacobian has $\rho(DF) = 5/4 > 1$, so the root is not locally attracting, although one start settles at it. The example establishes possibility only; trained-model prevalence and the validation of a check battery (Section 13.10) remain pending. A root solver's success is likewise not attraction, as Chapter 0 states (Section 0.4.3). *Pending:* Root solvers can recover repelling roots that none of the sampled starts reaches within the tested budget: root-solver success, residual and the attraction of the returned root, the spectral radius of the update's Jacobian there, by solver and system class.

13.5 Finite computation can change outer stability

Suppose another loop uses the refined state to update its input (Bruce, 2026h). If $T_K(A)$ comprises K inner updates of (1.4) from $X_0 = g(A)$, then

$$H_0 = Dg(A), \qquad H_{t+1} = DF_t H_t + \beta I, \qquad DF_t = D_X F_A(X_t), \quad DT_K(A) = H_K. \tag{13.1}$$

This derivative follows the whole trajectory, whereas the response $H = D_A X^*$ of Section 12.1 differentiates its equilibrium (Grazzi et al., 2020; Scieur et al., 2022). Input initialization gives $H_0 = I$; an input-independent saved state gives $H_0 = 0$. Fix a reference anchor A_0 and use the same inner initialization rule and budget for both terms in

$$G_K(A) = (1 - \zeta)A + \zeta\left[A_0 + C_\delta\left(\gamma[T_K(A) - T_K(A_0)]\right)\right], \tag{13.2}$$

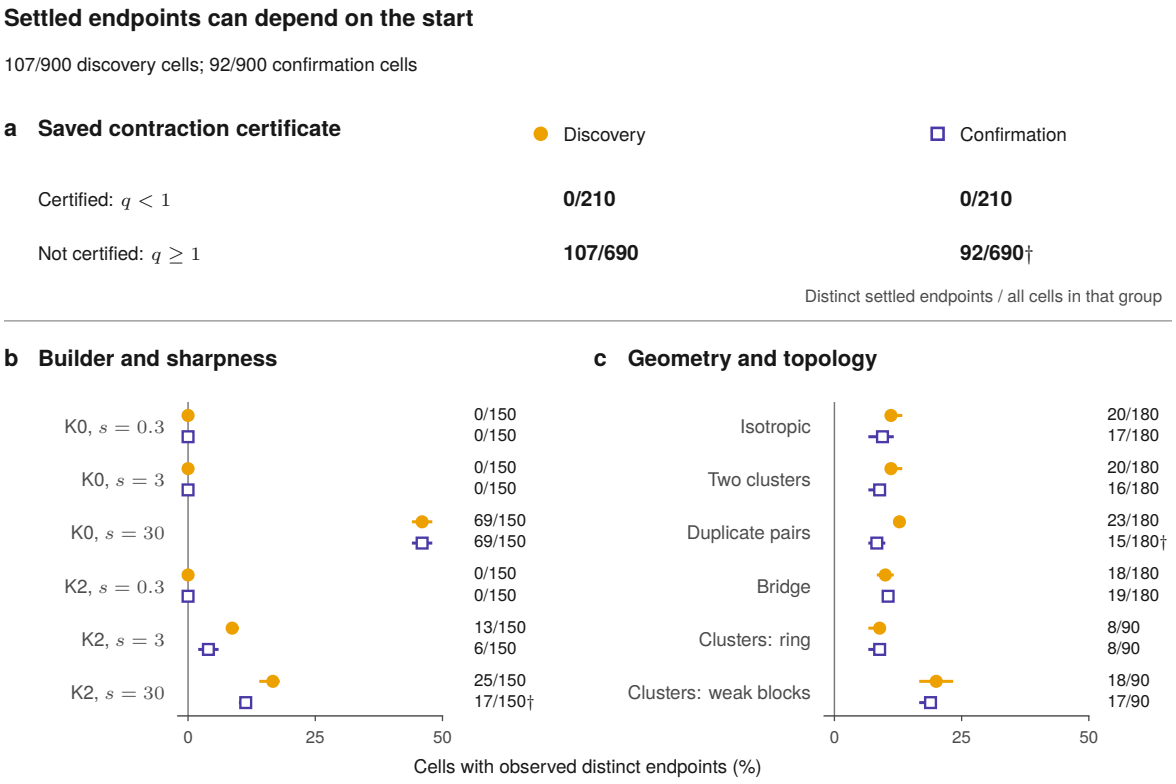


Figure 13.2: **Distinct settled endpoints were observed only outside the saved contraction certificate in this designed survey.** Starts at the input anchor and at zero are compared; the endpoint counts are those of Section 12.2. **(a)** Positive endpoint flags divided by all tested cells in each certificate class. The sufficient bound is $q = \alpha(1 + RL_P) < 1$, where α is the self-mixing coefficient, $R = 1.25$ the state-domain radius in maximum row Euclidean norm, and L_P the analytic Lipschitz upper bound for the mixing builder. **(b)** K0 uses dot-product softmax logits; K2 uses squared-distance softmax logits with a state-dependent bandwidth; s is logit sharpness. **(c)** The first four geometries use full connectivity and both state shapes $(N, d) = (16, 8), (64, 16)$ (N rows, d features); the last two use $(64, 16)$ two-cluster states with ring or weak-block topology. Circles (discovery) and open squares (confirmation) show the pooled fraction; horizontal segments span the minimum to maximum fraction across the three source seeds (0–2 and 100–102, respectively). Printed counts are descriptive summaries of the complete grid, not prevalence estimates or confidence intervals. Each phase has 900 cells over both builders, $s \in \{0.3, 3, 30\}$ and $\alpha \in \{0, 0.3, 0.6, 0.85, 0.95\}$; cells sharing a source seed are not independent replicates. A positive flag requires both starts to be labelled settled and their endpoint distance to exceed 10^{-4} in maximum row Euclidean norm. Settling requires relative residual $\leq 10^{-10}$ and maximum movement over 100 steps $\leq 10^{-9}$, each normalized by the larger of one and the current state norm. All 1,800 anchor starts settled. † One confirmation zero start (K2, $s = 30$, $\alpha = 0.95$, duplicate pairs, $(64, 16)$, seed 102) remained unresolved at 4,096 steps: its endpoint comparison is unknown and remains in every relevant denominator.

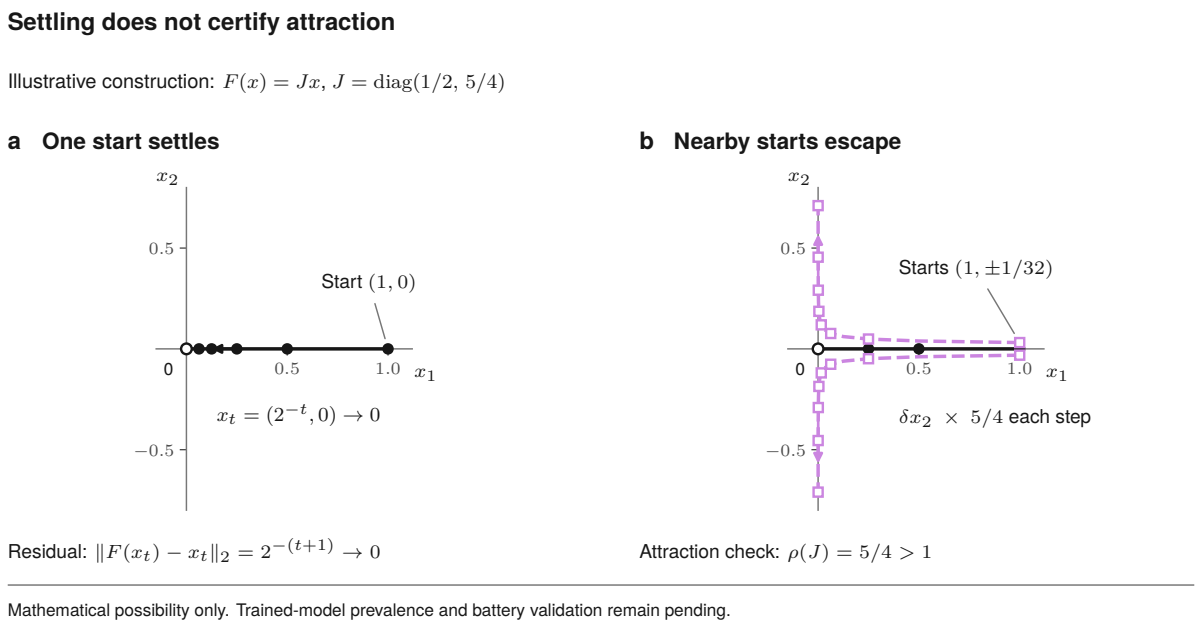


Figure 13.3: **A trajectory can settle at a root that is not locally attracting.** Illustrative construction, not measured model data: $F(x) = Jx$ with $J = \mathrm{diag}(1/2, 5/4)$ and root 0; the figure's J is the update's Jacobian, which the text writes DF. **(a)** From $(1, 0)$, $x_t = (2^{-t}, 0)$ and the residual tends to zero. **(b)** From $(1, \pm 1/32)$, the transverse displacement is $\pm(5/4)^t/32$ and grows at every step; the black orbit is repeated for comparison. Lines connect exact iterates through $t = 14$; filled circles and open squares mark selected iterates, and the open circle marks the root. Both panels use equal Euclidean axis scales. Here the original-map Jacobian has $\rho(J) = 5/4 > 1$, so the root is not locally attracting. This example establishes possibility only; trained-model prevalence and check-battery validation remain pending.

where $0 < \zeta \leq 1$ is damping, $\gamma \geq 0$ is feedback strength, and C_δ projects each row onto the Euclidean ball of radius δ. The subtraction fixes A_0 for every budget: changing K changes its local response without moving the reference point. Clipping bounds displacement and is inactive near A_0, where

$$DG_K(A_0) = (1 - \zeta)I + \zeta\gamma H_K. \tag{13.3}$$

Iterating (13.1) on a common contraction domain bounds the finite derivative at every budget, and with it the outer loop (Bruce, 2026h).

Corollary 13.1 (One outer condition for every budget)**.** *Under the common-domain contraction assumptions of Theorem 7.1, with $m > 0$ and $q = 1 - m$, and wherever the derivatives in (13.1) exist, the finite derivative satisfies, in the norm induced by the maximum-row norm,*

$$\|H_K\| \leq q^K\|H_0\| + \frac{\beta}{m}(1 - q^K). \tag{13.4}$$

Suppose $\|H_0\| \leq 1$. Since $0 < m \leq \beta$, this bound is a convex combination of $\|H_0\|$ and β/m, so $\|H_K\| \leq \beta/m$ for every $K \geq 0$. Consequently

$$\boxed{\gamma < m/\beta} \tag{13.5}$$

suffices for outer contraction at every budget, for input initialization and for any fixed detached initializer, on a common invariant domain.

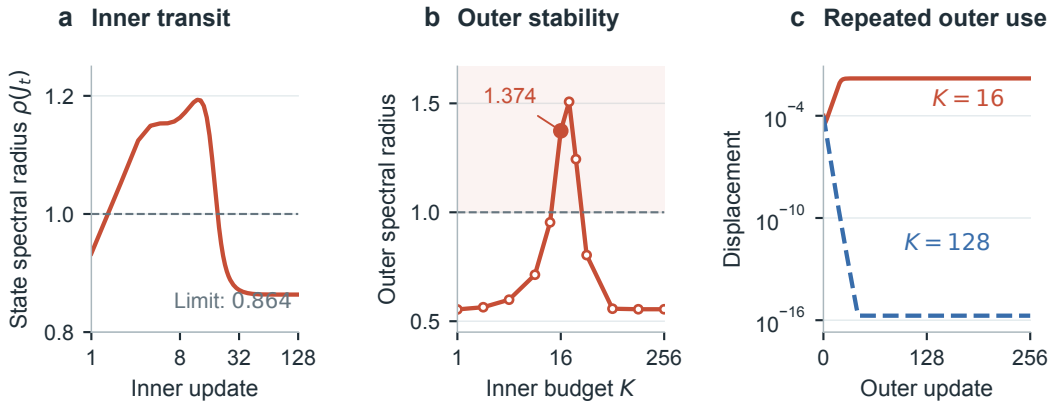


Figure 13.4: **An intermediate inner budget destabilizes the outer loop.** Duplicate-pair confirmation example with the Gaussian builder of adaptive bandwidth (K2), $\alpha = 0.85$, $s = 3$, $N = 16$, $d = 8$. (a) The inner trajectory passes through states with local spectral radius above one before reaching an attracting equilibrium, $\rho(DF) = 0.864$; the figure's J_t is the update's Jacobian DF_t. (b) With outer feedback $\gamma = 0.1$ and damping $\zeta = 0.5$, the full outer derivative has spectral radius 1.374 at $K = 16$ but falls below one at larger budgets. The horizontal boundary is one. (c) At the same $\gamma = 0.1$, the saved outer path grows for $K = 16$ and decays for $K = 128$. The budget sweep and spectral explanation are exploratory; the two outer paths were prespecified.

The proof is in Section A.14. Here β/m is the equilibrium-response bound of Corollary 7.2, and rowwise projection is nonexpansive in the same norm. The bound applies to a fixed number of updates; it does not differentiate the stopping decision of an adaptive solver. Outside this certificate, the actual finite derivative determines local stability.

Outside the certificate, an intermediate budget can destabilize the outer loop (Figure 13.4) (Bruce, 2026h). The controlled example uses input initialization on a duplicate-pair cloud of the designed geometry study with the Gaussian builder of adaptive bandwidth (K2; Section 12.2). It has a stable inner equilibrium, at which the update's Jacobian has spectral radius 0.864, yet $K = 16$, $\gamma = 0.1$ and $\zeta = 0.5$ give outer spectral radius 1.374. The first sampled perturbation shrinks by a factor of 0.512; repeated outer updates instead grow its displacement from 10^{-4} to 0.01559. One shrinking direction therefore misses an expanding mode. At $\gamma = 0.5$, the same short solve has spectral radius 4.869, while $K = 128$ gives 0.775 and decay. Central differences at amplitude 10^{-6} check the outer derivative along its dominant eigendirection at $K = 16$ for $\gamma = 0.1, 0.5$, with relative error below 1.2×10^{-9}. The two outer paths were prespecified; the budget sweep and the spectral explanation are exploratory.

Transient amplification occurs when DF_t expands before the trajectory settles (Bruce, 2026h). In the duplicate-pair trajectory, the local state spectral radius exceeds one during the transient before returning to its stable equilibrium value. The largest real eigenvalue of H_K grows from 1.083 at $K = 1$ to 17.476 at $K = 16$, is about 20.15 at $K = 20$, just past its largest value, 20.58 at $K = 19$, and falls to 1.099 at $K = 128$. The update length also increases during part of this transit. These are full derivatives along the saved nominal computation, not eigenvalues of one fixed linear map; their ordered products (Section A.14) determine the response. A scalar example makes the same distinction explicit (Figure 13.5). With $E = \mathbf{1}\mathbf{1}^\top/2$ and $Q = I - E$ as in Section 12.1, let $\alpha = 0.85$,

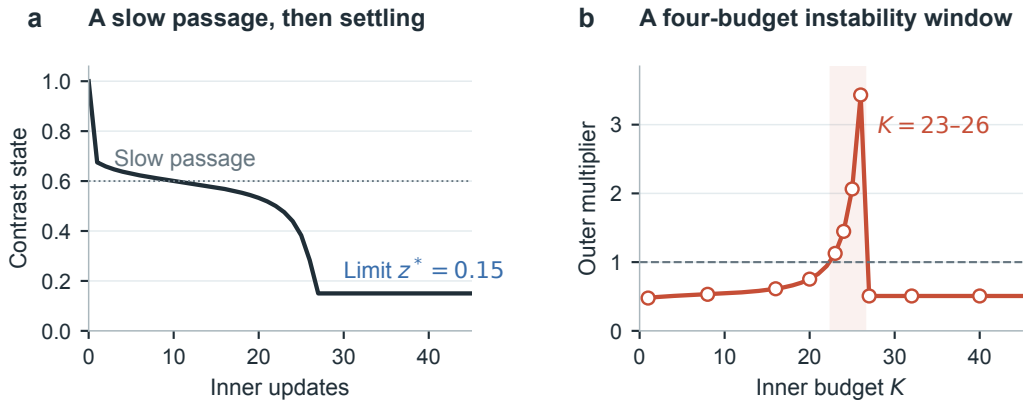


Figure 13.5: **Slow passage creates an intermediate instability window.** The explicitly specified scalar positive mixer converges to $z^* = 0.15$. (a) Its state trajectory passes slowly near 0.6 before reaching the constant branch. (b) The centred outer multiplier exceeds one only at budgets $K = 23$–26 in $1 \le K \le 128$, despite an equilibrium response of 0.15. Open circles are independent central differences.

$\beta = 0.15$, $A_0 = 1$, and

$$g(z) = \begin{cases} 0, & z \le 0.1, \\ \operatorname{clip}_{[0,1]}\left(\dfrac{z - 0.005 - 2(z-0.6)^2 - 0.15}{0.85z}\right), & z > 0.1, \end{cases} \qquad z^+ = 0.85g(z)z + 0.15A.$$

The two formulas agree on a neighbourhood of 0.1. The builder $E + g((x_1 - x_2)/2)Q$ is continuous and piecewise smooth, with nonnegative stochastic weights; the scalar map is its contrast update for anchors $(A, -A)$. Starting at $z_0 = A_0 = 1$, the trajectory remains in the invariant interval $[0, 1]$ and avoids the nonsmooth corners. On an unclipped segment it obeys $z^+ = z - 0.005 - 2(z - 0.6)^2$, with slow passage near 0.6 followed by faster motion. It eventually enters the constant segment $z^+ = 0.15$, whose multiplier is zero and equilibrium response is 0.15. Differentiating the exact finite recurrence gives a peak response of 58.6 at $K = 26$. For the centred outer map with $\zeta = 0.5$ and $\gamma = 0.1$, the multiplier $0.5 + 0.05H_K$ exceeds one exactly at $K = 23, 24, 25, 26$ in the checked range $1 \le K \le 128$; after entering the constant segment it is 0.5075 for all subsequent budgets. Central differences with step 10^{-7} agree within 10^{-5}. These examples lie outside the global contraction certificate; within it, (13.5) controls every budget. The spectral explanation and the scalar construction are exploratory. *Pending:* The response of a K-step computation: the finite-horizon response norm and the directional gain against the budget K, with the equilibrium response and the outer spectral radius, for the duplicate-pair example of this section and for a bank of systems.

13.6 Outer loops: directional probes miss expanding modes

Chapter 0 states that uniformly random unit probes in n dimensions read a rank-one gain G as $G/\sqrt{n}$ in root mean square, so that a directional probe can under-read an outer loop's gain (Section 0.4.3). The proof is in Section A.14; the full treatment, with clipping that can hide an expanding mode, is left for later work. Section 13.5 has shown one case: a first sampled perturbation

that shrinks by a factor of 0.512 while the outer spectral radius is 1.374, so that one shrinking direction misses an expanding mode.

13.7 What EMA moves and what it does not

An exponential moving average (EMA) can be put on the operator. In the constructions of Figure 13.6, the operator is built from an averaged state, $X_{t+1} = \alpha\mathcal{P}(\bar{X}_t)X_t + \beta A$ with $\bar{X}_{t+1} = (1-\eta)\bar{X}_t + \eta X_{t+1}$, so that $\eta = 1$ is the iteration without averaging. In these constructions, operator EMA settles oscillations while most predicted live-state fluctuation remains. All 54 saved outcomes, at averaged-state update rates $\eta \in \{1, 0.1, 0.01\}$ for 12 cloud systems, three descending fixtures and three signed witnesses, end in observed settling, an observed period two, or one of five horizons of 4,096 updates left unresolved, which are not divergence; the period-two outcomes are the descending fixtures at $\eta = 1$, which settle at both smaller rates. Separately, saved post hoc linear predictions for white forcing of amplitude 10^{-4} at the operator's builder input compare each system's variance at $\eta = 0.01$ with its variance at $\eta = 1$, for the live state, the averaged state and the two combined. The live and averaged blocks were not separately measured in the nonlinear runs, and at $\eta = 1$ the combined trace counts the same fluctuations twice. These designed systems do not establish trained-model prevalence.

The trained counterpart is a weight-tied looped block (Table 13.1): one Pre-LN causal attention-plus-MLP block on multi-query associative recall, trained at 4, 16 and 64 loops with no anchor, with the input re-injected at every loop, with an EMA of the latent whose state starts at the input, or with both. This EMA acts on the latent, not on the operator as in Figure 13.6. Section 5.2 reports the re-injection contrast at 96 loops (Table 5.2); the per-seed table adds the horizons, the drift and the runs that did not train. Without either anchor, the horizon $K_{1/2}$, the first loop count $K \geq K_{\text{train}}$ at which accuracy falls to half its value at the training count K_{train}, is 20–21 loops when trained at 4 and 60–72 when trained at 16. With re-injection, and with both anchors, every run that trained is right-censored past 96 loops; with re-injection alone its batch-level drift ratio at 96 loops is 133.1–145.5. With the EMA alone, the drift ratio is 6.0–11.7 in the runs that trained, and the horizon is 35–44 loops when trained at 4 and past 96 when trained at 16. The drift ratio is the Frobenius distance of the latent from the input embedding relative to the embedding's norm, one batch-level ratio on the last 32 sequences, not per token. At 64 loops, the runs that trained number 0/6 without the EMA and 3/6 with it, and at 16 loops one re-injection run did not train.

13.8 Two loops that are safe alone can explode together

Two training loops that share state can each pass a single-loop certificate while the pair diverges, and a two-loop eigenvalue pencil certifies the pair and names the coupling to cut, so far only on synthetic rigs (Section 0.4); on a trained self-distillation stack, a single composite margin did not certify collapse (Section 13.8). Figure 13.7 shows both on an exact construction on $\mathbb{R}^2$: $x_{t+1} = DF(g)\, x_t$, with $DF(g) = \left(\begin{smallmatrix} 0.8 & g \\ g & 0.8 \end{smallmatrix}\right)$ and cross-gain $g \geq 0$. Each isolated loop contracts with factor 0.8 and has margin $m_i = 1 - 0.8 = 0.2$. The comparison bound is attained, with $\beta_i = 0.2$, $K_{ii} = 0$ and $K_{12} = K_{21} = g$, giving the pencil $H(g) = \left(\begin{smallmatrix} 0.2 & -g \\ -g & 0.2 \end{smallmatrix}\right)$. Its minimum eigenvalue is $0.2 - g$, and the common-mode multiplier is exactly $\rho(DF) = 0.8 + g$: a cross-gain of 0.1 contracts ($\rho = 0.9$), 0.2 is neutral, and 0.3 grows ($\rho = 1.1$). Reducing both cross-gains from 0.3 to 0.1 restores contraction. For general two-loop systems the certificate requires valid upper gain bounds; failure of a sign-blind bound need not imply instability of a signed system. Beyond two loops the symmetric pencil

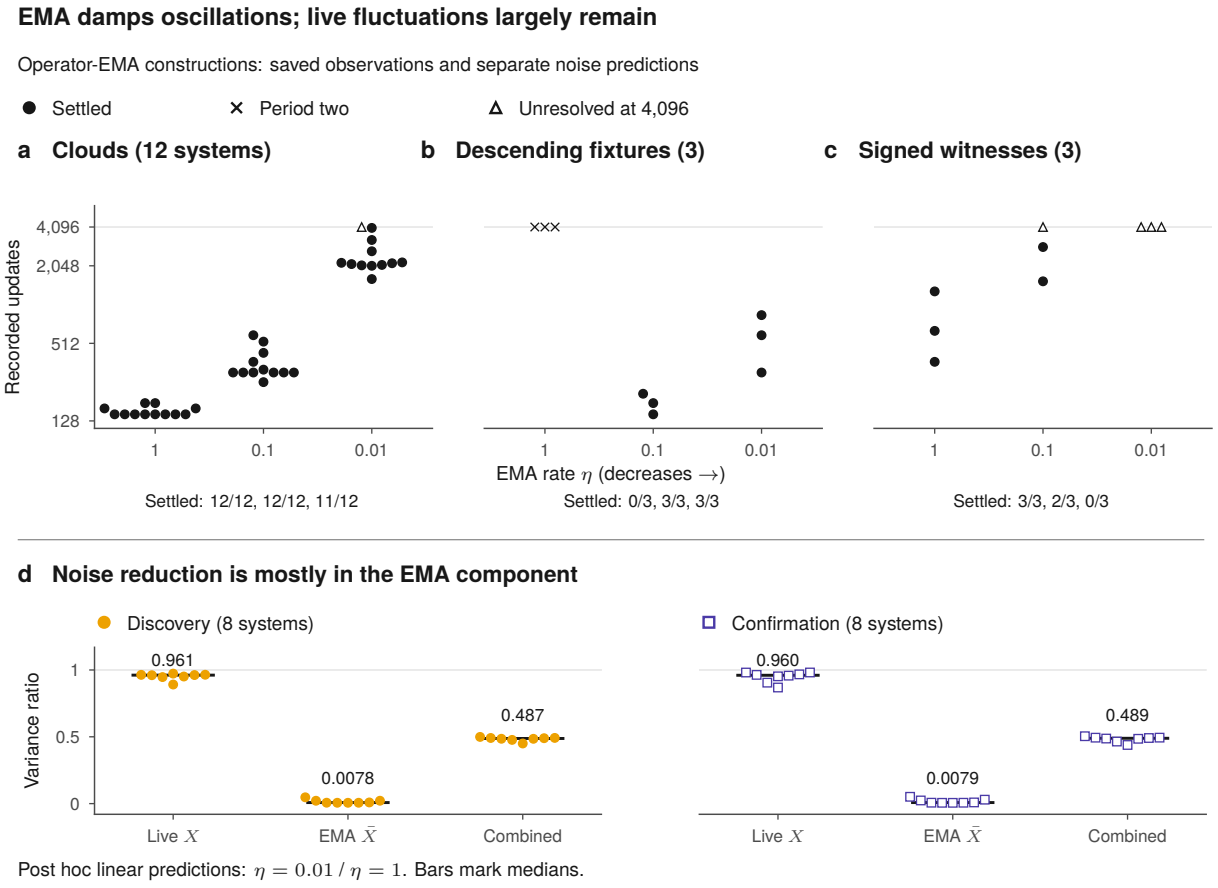


Figure 13.6: **In these constructions, operator EMA settles oscillations while most predicted live-state fluctuation remains. (a–c)** All 54 saved outcomes at averaged-state update rates $\eta \in \{1, 0.1, 0.01\}$: 12 cloud systems (source seeds 0–11), three descending fixtures and three signed witnesses (seed 0). Vertical position is the recorded update count on a shared logarithmic scale; horizontal offsets separate points within each rate. Filled circles denote observed settling, crosses observed period two, and open triangles the five unresolved 4,096-update horizons, not divergence. Counts follow the displayed rate order. **(d)** Saved post hoc linear predictions for white forcing of amplitude 10^{-4} at the operator's builder input. Each phase has eight fixed systems (four geometries by two builders). For each component, points show each system's variance at $\eta = 0.01$ divided by its variance at $\eta = 1$; bars and numbers give the median of those eight ratios. Each saved prediction averages eight forcing-direction replicates; all between-system spread is shown, with no inferential intervals. The live X and averaged $\bar{X}$ blocks were not separately measured in the nonlinear runs. At $\eta = 1$, $X = \bar{X}$, so the combined trace counts the same fluctuations twice. These designed operator-EMA systems do not establish trained-model prevalence or characterize the moving average of the latent in the trained looped block of Table 13.1.

Arm	K_{train}	Seed	a_1	a_2	a_4	a_8	a_{16}	a_{32}	a_{64}	a_{96}	$K_{1/2}$ [bracket]	D_{96}	Trained
none	4	0	0.083	0.997	1.000	0.998	0.719	0.155	0.049	0.032	21 [16, 32]	45.9	yes
none	4	1	0.075	1.000	1.000	0.999	0.713	0.152	0.046	0.031	21 [16, 32]	46.9	yes
none	4	2	0.101	0.998	1.000	1.000	0.666	0.103	0.033	0.023	20 [16, 32]	54.6	yes
none	16	0	0.030	0.105	0.722	0.997	1.000	0.998	0.621	0.228	72 [64, 96]	46.8	yes
none	16	1	0.027	0.077	0.702	0.990	1.000	0.987	0.444	0.183	60 [32, 64]	57.6	yes
none	16	2	0.026	0.071	0.402	0.939	0.990	0.953	0.522	0.225	66 [64, 96]	46.9	yes
none	*64*	*0*	*0.016*	*0.021*	*0.048*	*0.107*	*0.139*	*0.160*	*0.172*	*0.161*	*did not train*	*50.6*	*no*
none	*64*	*1*	*0.014*	*0.026*	*0.054*	*0.101*	*0.140*	*0.154*	*0.166*	*0.165*	*did not train*	*69.7*	*no*
none	*64*	*2*	*0.016*	*0.019*	*0.050*	*0.098*	*0.138*	*0.154*	*0.172*	*0.166*	*did not train*	*55.7*	*no*
reinj	4	0	0.123	0.998	1.000	1.000	0.998	0.972	0.933	0.904	> 96	137.4	yes
reinj	4	1	0.129	0.999	1.000	1.000	0.999	0.990	0.972	0.959	> 96	136.7	yes
reinj	4	2	0.155	0.995	1.000	1.000	0.997	0.980	0.954	0.939	> 96	145.5	yes
reinj	*16*	*0*	*0.133*	*0.160*	*0.184*	*0.199*	*0.207*	*0.206*	*0.189*	*0.184*	*did not train*	*191.9*	*no*
reinj	16	1	0.123	0.430	0.976	1.000	1.000	1.000	0.999	0.997	> 96	133.1	yes
reinj	16	2	0.145	0.400	0.953	0.996	0.999	0.999	0.997	0.994	> 96	136.7	yes
reinj	*64*	*0*	*0.016*	*0.018*	*0.015*	*0.021*	*0.030*	*0.028*	*0.029*	*0.031*	*did not train*	*1175.2*	*no*
reinj	*64*	*1*	*0.017*	*0.021*	*0.020*	*0.017*	*0.016*	*0.016*	*0.016*	*0.016*	*did not train*	*1490.3*	*no*
reinj	*64*	*2*	*0.018*	*0.018*	*0.014*	*0.016*	*0.016*	*0.015*	*0.015*	*0.015*	*did not train*	*1520.5*	*no*
ema	4	0	0.130	0.980	1.000	0.999	0.951	0.548	0.156	0.077	35 [32, 64]	11.3	yes
ema	4	1	0.116	0.991	1.000	1.000	0.963	0.693	0.270	0.134	44 [32, 64]	10.5	yes
ema	4	2	0.137	0.986	1.000	0.997	0.941	0.560	0.171	0.087	36 [32, 64]	11.7	yes
ema	16	0	0.085	0.391	0.953	1.000	1.000	1.000	0.986	0.884	> 96	6.0	yes
ema	16	1	0.063	0.471	0.950	1.000	1.000	1.000	0.983	0.843	> 96	7.2	yes
ema	16	2	0.091	0.388	0.946	1.000	1.000	1.000	0.989	0.835	> 96	8.2	yes
ema	64	0	0.031	0.055	0.187	0.685	0.973	1.000	1.000	1.000	> 96	7.7	yes
ema	64	1	0.025	0.035	0.144	0.606	0.943	0.998	1.000	0.999	> 96	9.9	yes
ema	*64*	*2*	*0.024*	*0.051*	*0.106*	*0.160*	*0.226*	*0.301*	*0.333*	*0.315*	*did not train*	*10.9*	*no*
ema+reinj	4	0	0.136	0.986	1.000	1.000	0.995	0.972	0.883	0.762	> 96	20.5	yes
ema+reinj	4	1	0.124	0.995	1.000	1.000	0.993	0.984	0.935	0.863	> 96	19.8	yes
ema+reinj	4	2	0.157	0.993	1.000	0.999	0.997	0.982	0.926	0.839	> 96	20.3	yes
ema+reinj	16	0	0.131	0.594	0.980	1.000	1.000	1.000	0.997	0.994	> 96	16.8	yes
ema+reinj	16	1	0.118	0.801	0.992	1.000	1.000	1.000	1.000	0.999	> 96	17.4	yes
ema+reinj	16	2	0.138	0.734	0.987	1.000	1.000	1.000	1.000	0.994	> 96	17.4	yes
ema+reinj	*64*	*0*	*0.080*	*0.158*	*0.184*	*0.214*	*0.231*	*0.237*	*0.247*	*0.233*	*did not train*	*20.8*	*no*
ema+reinj	*64*	*1*	*0.019*	*0.022*	*0.021*	*0.019*	*0.018*	*0.015*	*0.015*	*0.014*	*did not train*	*181.1*	*no*
ema+reinj	64	2	0.121	0.322	0.661	0.889	0.972	0.993	0.996	0.995	> 96	17.2	yes

Table 13.1: **Depth extrapolation of a weight-tied looped block, per seed.** One weight-tied Pre-LN causal attention-plus-MLP block on multi-query associative recall, trained at $K_{\text{train}} = 4$, 16 and 64 loops with neither input re-injection nor averaging (none), with the input re-injected at every loop (reinj), with an exponential moving average of the latent that starts at the input (ema), or with both (ema+reinj); a_K is accuracy after K loops, and Table 5.2 prints the first three arms at 4 and 16 loops from the same records. 2,048 answers per value (256 sequences); chance 1/64. Conservative sequence-level 95% interval half-width is at most 0.063. $K_{1/2}$ is the first loop count $K \geq K_{\text{train}}$ at which accuracy falls to half its value at K_{train}, interpolated linearly in $\log K$. Horizons are rounded to the nearest loop with their bracketing grid points; > 96 means right-censored. Untrained cells are italicized. Drift at 96 loops, D_{96}, is one batch-level Frobenius ratio on the last 32 sequences, not per token.

requires balanced gains; the general certificate is $\rho(D_\beta^{-1}K) < 1$, with $D_\beta = \text{diag}(\beta_i)$. The proof of the certificate is in Section A.14. The construction is a mathematical illustration, with no empirical uncertainty or trained-model measurements.

A single composite margin is not enough, even when the direction it watches is tracked faithfully. On a trained self-distillation stack, in which a student learns from a moving-average teacher while a labelled task anchors it, the composite margin m_{eff} is the average of two quantities: the margin of the self-distillation loop alone, and the curvature that the task supplies along the measured worst direction. The theory under test predicts collapse exactly when this margin is negative. Of 25 runs on CIFAR-10, five fractions of labelled data with five seeds each, trained for 200 epochs, four

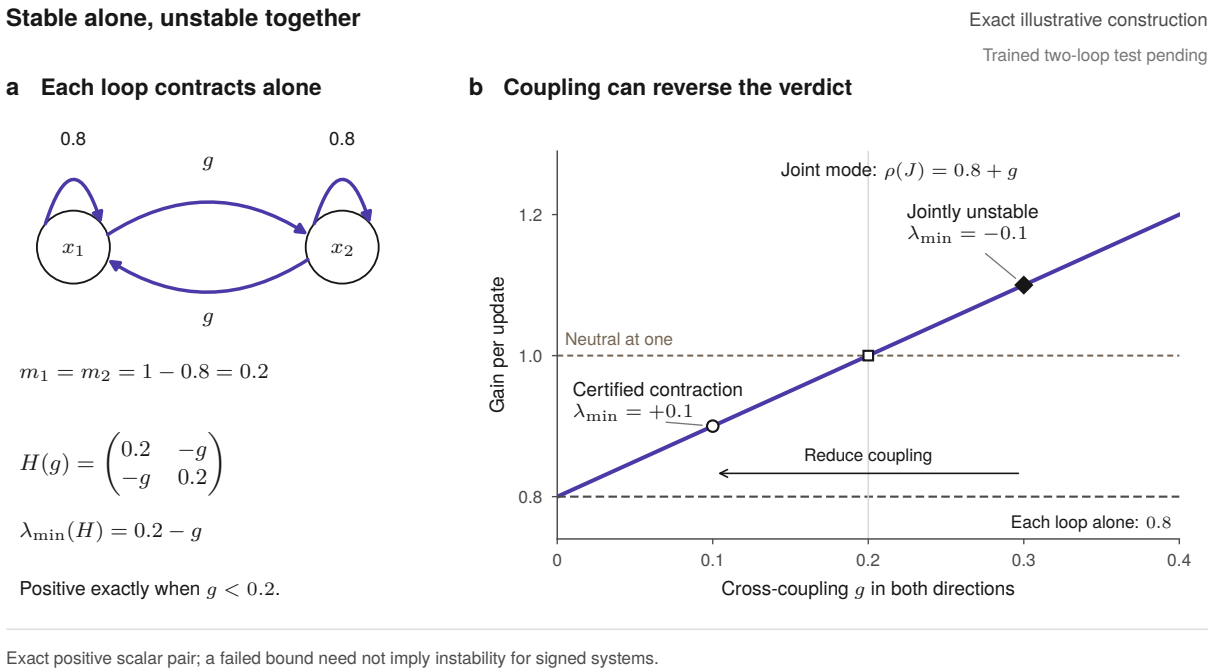


Figure 13.7: **Individually contracting loops can become jointly unstable when cross-coupling exceeds their margin. (a)** Exact illustrative construction on $\mathbb{R}^2$: $x_{t+1} = J(g)x_t$, with $J(g) = \left(\begin{smallmatrix} 0.8 & g \\ g & 0.8 \end{smallmatrix}\right)$, $g \geq 0$; the figure's J is the joint update's Jacobian, which the text writes DF. Each isolated loop has margin $m_i = 0.2$. Arrow labels are exact update coefficients; arrow widths are schematic. The comparison bound is attained, with $\beta_i = 0.2$, $K_{ii} = 0$, and $K_{12} = K_{21} = g$, giving the displayed pencil $H(g)$. **(b)** Its minimum eigenvalue is $0.2 - g$, and the common-mode multiplier is exactly $\rho(J) = 0.8 + g$. The open circle at $g = 0.1$ contracts ($\rho = 0.9$); the square at $g = 0.2$ is neutral; the diamond at $g = 0.3$ grows ($\rho = 1.1$). Reducing both cross-gains from 0.3 to 0.1 restores contraction. For general two-loop systems the certificate requires valid upper gain bounds; failure of a sign-blind bound need not imply instability of a signed system. Beyond two loops the symmetric pencil requires balanced gains; the general certificate is $\rho(D_\beta^{-1}K) < 1$. This is a mathematical illustration, with no empirical uncertainty or trained-model measurements; the trained two-loop test is pending (Section 13.9).

collapsed, each through a collapse of the teacher's representation that its effective rank detected. Two collapsed early, at label fractions 0.3 and 1, at epoch index 16 in each run, counting from zero, although their lowest measured composite margins were approximately 8.53 and 16.31, respectively. Two collapsed late, at fractions 0.01 and 0.03, at epoch indices 106 and 142, in runs whose lowest margins were -1.64 and -0.46. In every collapsed run the worst direction was tracked, with a median cosine of at least 0.99 between its successive estimates. As a classifier of which runs collapse, the composite margin has an area under the curve of 0.44, at chance, against 0.62 for the self-distillation loop's measured gain alone; with four collapses, fewer than the ten per class that this comparison needs, it is underpowered and no evidence that the margin orders the runs backwards. The failure is nonetheless real, and scoped. The test fixed in advance attached one condition to it, that the worst direction be tracked, and that condition held: a scalar can follow its direction faithfully and still watch the wrong subspace, for the margin follows the curvature of the task head

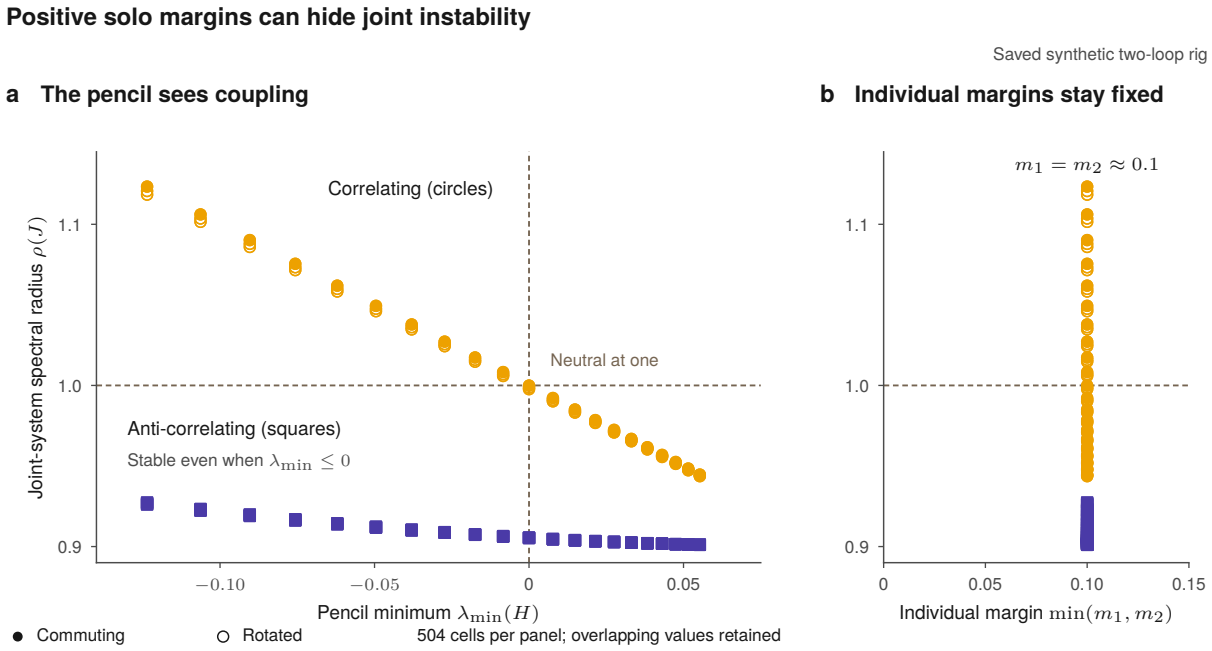


Figure 13.8: **Positive individual margins do not certify the coupled system, while the two-loop pencil exposes the correlating instability and remains conservative for opposing couplings.** (a) The spectral radius $\rho(J) = \max_j |\lambda_j(J)|$ of each saved joint-system matrix versus the measured pencil minimum $\lambda_{\min}(H)$; the figure's J is the joint update's Jacobian DF. Here $H_{ii} = m_i = 1 - \widehat{K}_{ii}$ and $H_{12} = H_{21} = -\sqrt{\widehat{K}_{12}\widehat{K}_{21}}$, using the stored power-ascent block gains. (b) The same radii versus the recorded individual comparator $\min(m_1, m_2)$, the best individual margin; both margins are pinned to 0.1 up to roundoff. Circles (amber) are correlating and squares (violet) anti-correlating; filled marks are the commuting arm and open marks the rotated arm. Each panel retains all 504 sweep cells: two update rates $\beta \in \{0.1, 0.3\}$, two modes, two sign quadrants, 21 coupling products and seeds 0–2. Raw seed observations overlap; no interval is newly estimated. Dashed lines mark $\rho(J) = 1$ and pencil zero. Of 252 anti-correlating cells, 124 have a non-positive pencil yet remain stable. Boundary roundoff and all stored outcomes are retained. The separation is not exactly sharp: for the commuting, correlating cells the area under the curve is 0.9998739, against the 1 that a sharp boundary requires, because of ties on the zero grid. These are saved synthetic two-loop results, not trained-loop evidence or a universal multi-loop certificate.

while the collapse happens in the teacher's representation. What fails is certification by one number. A retrospective fit still relates the margin to collapse time: over the four collapses, the time to collapse scales as a power of the lowest margin's magnitude, with exponent -0.70 ($R^2 = 0.92$), at the edge of the band from -1.3 to -0.7 fixed in advance. The remedy that the test names for this failure is a certificate that resolves the two directions, the two-loop pencil, which the rest of this section tests on synthetic rigs. The sweep covers one dataset, and the pencil on trained loops is still to be tested (Section 13.9).

On a synthetic two-loop rig, positive individual margins do not certify the coupled system, while the two-loop pencil exposes the correlating instability and remains conservative for opposing

couplings (Figure 13.8). Across all 504 sweep cells both loops' margins are pinned at 0.1, up to roundoff, so the individual margins cannot tell the stable cells from the unstable ones; the pencil, built from the stored block gains, crosses zero where the spectral radius of the commuting, correlating cells crosses one. Of the 252 anti-correlating cells, 124 have a non-positive pencil yet remain stable: there the pencil is conservative. The separation is not exactly sharp: for the commuting, correlating cells the area under the curve is 0.9998739, against the 1 that a sharp boundary requires, because of ties on the zero grid. These are saved synthetic two-loop results, not trained-loop evidence or a universal multi-loop certificate.

Where is the certificate tight? Across all 630 production two-loop cells, with 32 dimensions per loop, the relative gap G between the true stability boundary and the pencil's falls below 0.05 in 77 correlating cells and in no anti-correlating cell (Figure 13.9). The gap is mapped against two coordinates of the coupled blocks, the Frobenius norm of a commutator and the Henrici departure from normality, and the near-tight maxima of these coordinates are marginal extents, not a guarantee for the enclosed rectangle or for unobserved points. A further 154 cells have a zero pencil boundary, so that their relative gap is undefined, although their true boundaries are positive. This is a descriptive map: the five-sigma separation of tight from conservative cells that was asked for in advance is not reached ($2.643 < 5$), and no trained-system or unbalanced multi-loop guarantee is inferred.

Acting on the certificate, a governor whose auxiliary coupling is set by the pencil can be compared with a naive controller in recursive training (Figure 13.10). In a saved synthetic comparison, the governor meets the proxy-quality tolerance fixed in advance at one quarter of the fixed-fraction budget, with zero degradation events in six of six lineages, while over-driving produces alternating drift. The quoted +4.02% is the stored upper endpoint of a 95% paired-bootstrap interval for the governor minus the fixed fraction, divided by the fixed fraction's mean, from three pair means. The proxy is computed by a saved scalar drift model, not a trained language model. An archived language-model reference, with WikiText-2 validation perplexities for a controller against a fixed fraction, is shown apart and not pooled. Detector firing, degradation and the scalar-multiplier divergence flag are distinct: the over-driven arm fires in six of six lineages with zero degradation events, while the no-loop control fires in one of six. This is a synthetic governor comparison, not a trained-governor result or a general signed-system certificate.

13.9 In trained systems

Pending: Trained looped transformers and implicit models: the settling residual, the resolved response, the finite-budget derivative against K and the attraction of the root, with and without input re-injection, and per-model rates of under-resolved responses, start-dependent endpoints, non-attracting roots and finite-budget peaks, with intervals.
Pending: A collapse observable for recursive training: lineage growth and validation perplexity against the retained fraction of real data, and the separation of candidate collapse observables against a frozen five-sigma gate.
Pending: The pencil in a trained two-loop system: collapse incidence against the measured pencil margin, the two per-loop margins and the scalar composite margin.

13.10 A check battery for fixed-point computations

Each reading of Section 13.1 has a check that this chapter has already shown. For the response, resolve the derivative at the precision it needs: independent 80-digit roots recover the gain that a

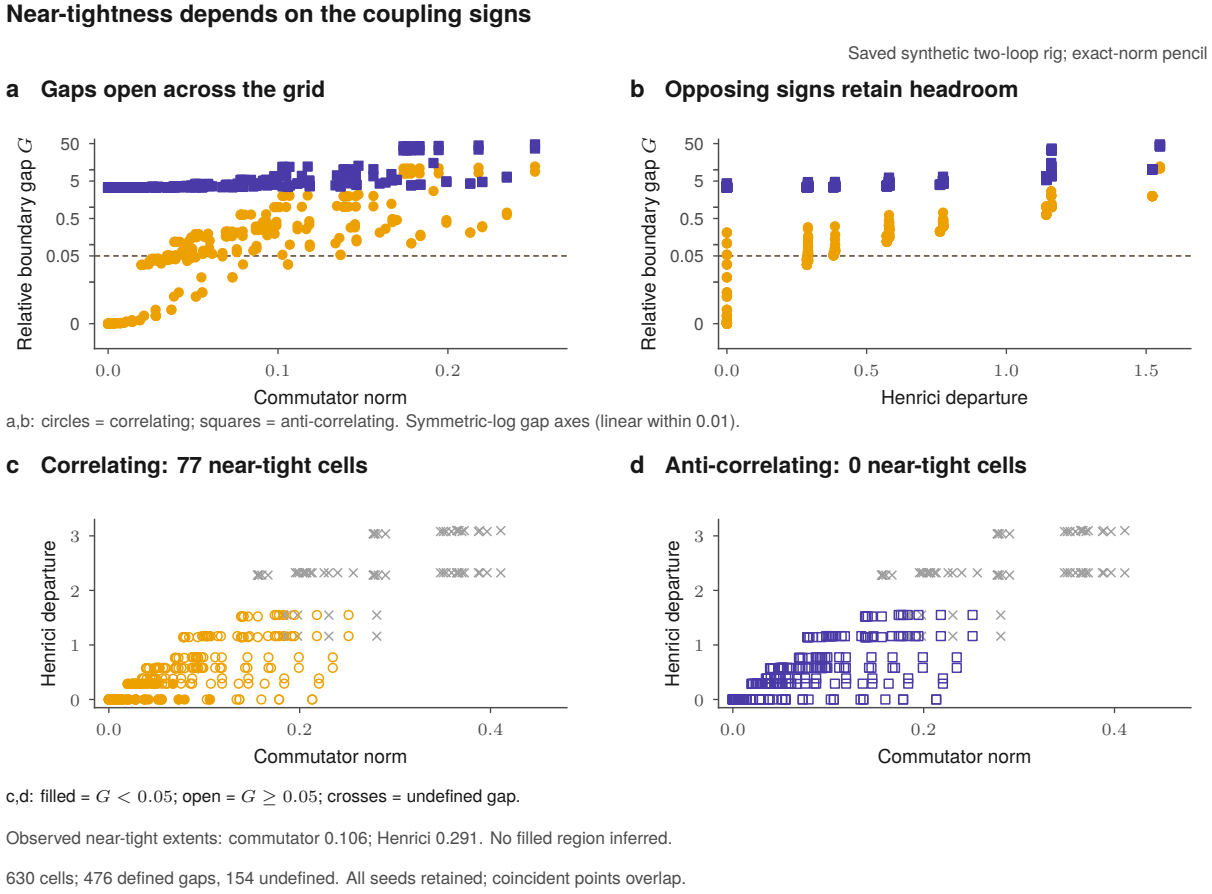


Figure 13.9: **Near-tight boundaries occur only in a measured subset of correlating cells; anti-correlating couplings retain conservative headroom.** (a,b) The saved relative gap $G = (\alpha^*_{\text{true}} - \alpha^*_{\text{pencil}})/\alpha^*_{\text{pencil}}$ against each realized coordinate, with correlating circles and anti-correlating squares. The pencil boundary uses exact block operator norms (true K); α^*_{true} is the saved boundary of $\rho(J(\alpha)) < 1$, from bisection to 10^{-6}. Here α scales the coupling, and the figure's J is the joint update's Jacobian DF. (c,d) All coordinate pairs, separated by sign quadrant. Filled marks identify the observed $G < 0.05$ region: 77 correlating cells and no anti-correlating cells; open marks have $G \geq 0.05$. Crosses retain the 154 cells with zero pencil boundary and undefined relative gap, excluded from (a,b). Their true boundaries are positive; a cross does not assert instability. Coordinates are the raw Frobenius commutator $\|[J_{22}(1), J_{21}(1)]\|_F$ and Henrici departure $[\|J_{11}\|_F^2 - \sum_j |\lambda_j(J_{11})|^2]^{1/2}$, read directly from the saved records. The reported near-tight coordinate maxima are marginal extents, not a guarantee for the enclosed rectangle or unobserved points. All 630 production two-loop cells are shown: seven rotations, five non-normality settings, two signs, three update-rate pairs and three structure seeds; 32 dimensions per loop. All raw seeds and numerical boundary residuals remain, with no fitting, jitter or estimated intervals; coincident points overlap. Every instrument check passes, but the five-sigma separation of tight from conservative cells that was asked for in advance is not reached ($2.643 < 5$); this is a descriptive map. No trained-system or unbalanced multi-loop guarantee is inferred.

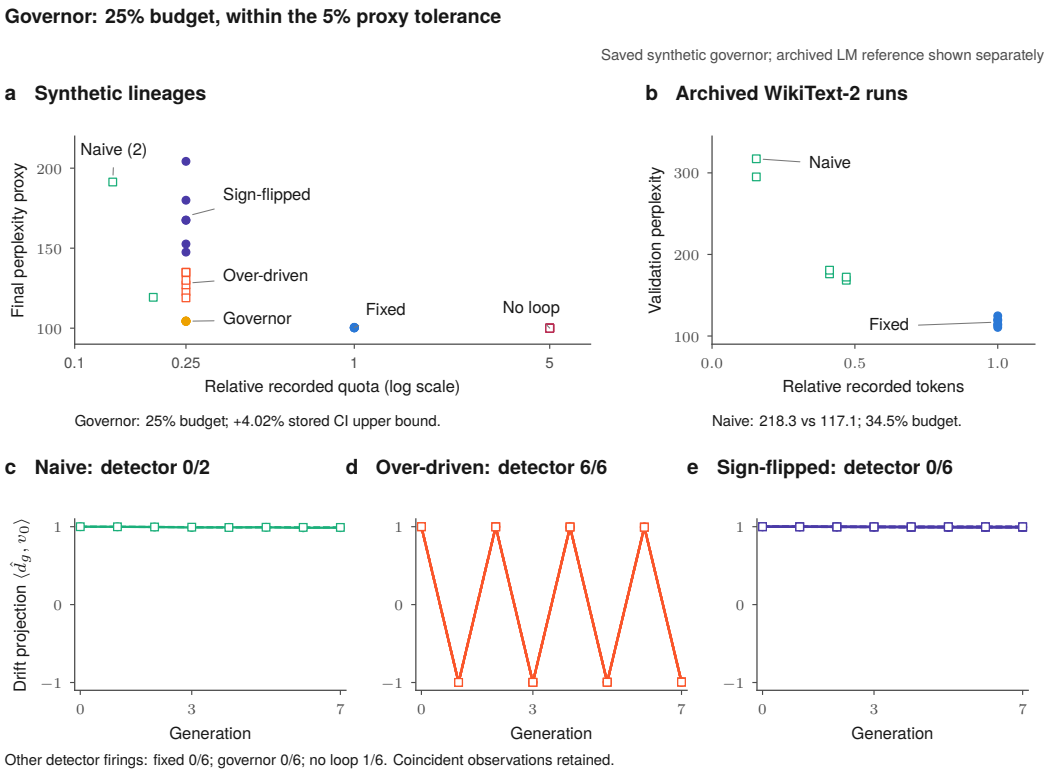


Figure 13.10: **The saved synthetic governor meets the proxy-quality tolerance fixed in advance at one-quarter of the fixed-fraction budget, while over-driving produces alternating drift. (a)** Final *perplexity proxy* of the synthetic lineages versus recorded real-data quota, normalized by the fixed arm's mean (logarithmic horizontal axis). All 32 policy and control lineages are shown: a fixed fraction ($\beta = 0.2$), a naive controller, the governor ($\beta = 0.05$), an over-driven arm, a matched-magnitude sign-flipped arm, and a no-loop control. Each arm has three paired seeds (six lineages), except the naive controller's one pair (two). The proxy is computed by the saved scalar drift model, not a trained language model; the quota meter is the saved $\sum_g \beta_g$. **(b)** Separate archived WikiText-2 validation perplexities of a language-model reference: controller versus fixed $\beta = 0.2$, three pairs (six lineages) each after twelve generations. Recorded tokens are normalized within that reference; its original nominal $g = 0$ accounting is retained, although that corpus is pure real data. Absolute quality is not pooled across the two experiments. **(c–e)** All eight saved projection samples $\langle \hat{d}_g, v_0 \rangle$ per synthetic lineage, relative to its initial direction; solid circles and dashed open squares identify the two lineages within each pair. Coincident observations overlap without jitter. The quoted +4.02% is the stored upper endpoint of a 95% paired-bootstrap interval for the governor minus the fixed fraction, divided by the fixed fraction's mean, using *three pair means*; no interval is re-estimated. The governor has zero degradation events in six of six lineages. Detector firing, degradation and the scalar-multiplier divergence flag are distinct: the over-driven arm fires in six of six with zero degradation events, while the no-loop control fires in one of six. The naive controller ends at $\beta = 0.0200$ and 0.0222, not both at the literal floor. By the criteria fixed in advance, the synthetic comparison succeeds and the archived reference does not. This is a synthetic governor comparison, not a trained-governor result or a general signed-system certificate.

float64 difference misses (Section 13.2). For uniqueness, settle from more than one start, at the input anchor and at zero, and compare the endpoints (Section 13.3). For attraction, compute the spectral radius of the update's Jacobian at the root that was found, not only the residual there (Section 13.4). For an outer loop, compute the outer derivative from the finite derivative H_K at the budget actually used, and its spectral radius, not the gain along one sampled direction (Section 13.5). The battery has not yet been validated on trained models (Section 13.9).

For anchored adaptive averaging, Section 7.8 gives a practical sequence of checks with Table 7.2 and its scope (Bruce, 2026h).

13.11 What holds in trained networks

Exact for every network of the stated class. Settling is not solving (Sections 13.1 to 13.10). A negative result that holds for any solver: settling certifies none of response, uniqueness, attraction or outer-loop stability; shown by explicit constructions. The results are proved and shown on constructions and small systems, and the coupled loops' two-loop certificate so far only on synthetic rigs (on a trained self-distillation stack a single composite margin did not certify collapse, Section 13.8); trained looped and implicit models are still to be tested (Section 13.9).

Part IV

Why the transformer sits where it does

Chapter 14

The lattice and the sweet spot

Parts I and II named what an averaging layer that is rebuilt from its input and iterated needs: an anchor, a free content path, gain control, a way to abstain, a routing budget that grows with length and an arrow, with a writer beside them. This part reads the standard block against that list. Chapter 14 asks which part of the block pays for each condition, and at what price. Chapter 15 asks why local changes to the block rarely pay. Chapter 16 asks whether designs jump rather than drift, which is a hypothesis under assumed concave prices; it then reads, retrospectively, how practice found the conditions, and assembles the smallest block that meets them, a reading of an assembled constraint set.

Chapter 0 states the sweet spot in brief (Section 0.3.3), and Figure 0.2 gives one part of the standard block to each need. That match holds by construction: the inventory was assembled by asking what each part of the block protects. What does not follow from the construction is each row's necessity within its class, which Part II proves, and the predictions about other designs, which Chapter 20 states. This chapter develops both the sweet spot and the figure from the lattice of constraints (Bruce, 2026d): what each constraint of the block guarantees, what relaxing it buys and what must then be re-supplied (Sections 14.1 and 14.2), with signed mixing as the case in point (Section 14.3); the prices that decide among designs (Section 14.4); and the sweet spot as a reading under those prices (Section 14.5). Section 14.6 says what of this holds in trained networks.

14.1 How the standard block pays for each condition

Tables 14.1 to 14.3 give the lattice (Bruce, 2026d). Each row is a constraint. The "vertex" marks where softmax attention in a pre-normalized residual block sits. The columns give what relaxing the constraint buys, what must then be re-supplied, and where practice sits, and the last column marks the rows with a result still pending. Figure 14.1 draws the lattice (Section 14.2). The inventory of conditions (Table 0.2) does not say that every successful sequence model implements six named mechanisms; it says what goes wrong, and at what price, when a model in the stated class lacks a property. Like the inventory, the rows are conditional: each is a statement about a declared class, and the notes after the tables give each row's scope and source.

The kernel constraints come first (Table 14.1). At the vertex, unit row sums and density are satisfied, positivity holds per head, and reversibility and the positive-semidefinite score form are relaxed.

Table 14.1: Kernel constraints.

Constraint (vertex)	Relax → buys	Must re-supply	In practice	Open
A1 Unit row sums (satisfied)	Abstention; strict contraction $\|P\|_w = \rho < 1$ for an irreducible head (Theorem 8.3); sink-free routing; definite logit Jacobian	Length calibration of the null, within the two-sided budget; norm control when normalization is dropped	Sinks, registers, off-by-one, learned sink logits, output gates, sigmoid $-\log n$ with hybrid norm	pending
A2 Positivity (per head)	Subtraction, contrast, high-pass filters, Perron deflation	Norm cap (spectral contraction) or per-head normalization	Signed W_O across heads; differential attention	pending
A3 Reversibility (relaxed)	Direction; rotation	Nothing in $L^2(\pi)$ (circulation rotates, never amplifies); the current on an edge is at most twice its reciprocal traffic	Untied Q/K; causal mask in decoders; RoPE	pending
A4 PSD score form (relaxed)	Negative eigenvalues (contrast inside reversible mixing)	Nothing new; still the anchor for the constant mode	Indefinite learned W_{QK}	—
A5 Density (satisfied)	Leak-free routing; per-block filters	Block discovery; gradients to excluded keys	Sliding window, document masks, sparse attention	—

The dynamical and context constraints follow (Table 14.2). At the vertex the kernel is adaptive rather than frozen, the key–value cache is unbounded, and the temperature is fixed.

Table 14.2: Dynamical and context constraints.

Constraint (vertex)	Relax → buys	Must re-supply	In practice	Open
B1 Frozen kernel (relaxed: adaptive)	Content-dependent routing	Gain control, a positive margin $\beta - \alpha\kappa_{\mathrm{f}}$ (with row mass: $q < 1$): anchor, bounded logit scale, bounded radius	Residual, LayerNorm, QK-norm, soft-capping (z-loss acts on the output logits)	pending
B2 Unbounded KV (vertex) vs. bounded contracting state	Length robustness; $O(1)$ state for a pure recurrence	Recall reserve (a few attention layers, whose cache stays linear in context)	SSMs, (gated) delta rules; hybrids at $1/k$ attention	pending

Constraint (vertex)	Relax → buys	Must re-supply	In practice	Open
B3 Fixed temperature (vertex) vs. $\log N$ budget	Retrieval and abstention that survive length	A null, and positions kept in distribution when extending; scaling every logit by $\log N / \log N_{\text{train}}$ helped where positions were interpolated, in vision transformers at higher resolution and in 125M causal language models, and raised long-context perplexity on top of YaRN, as did YaRN's temperature alone without interpolation	$\log n$ scaling, YaRN temperature, scalable softmax	pending

The block protections close the lattice (Table 14.3): the anchor, alignment, norm and gain control, write capacity and head multiplicity. For each, the columns give the signature of its removal and whether capacity elsewhere can compensate for it, which Chapter 15 takes up.

Table 14.3: Block protections.

Protection (vertex)	Remove → signature	Compensable?	In practice	Open
C1 Anchor (Pre-LN residual; critical at initialization)	Tenfold loss at $\approx 2.3/(1-\mu_C)$ mixing steps; a fixed share p only delays it, to $\approx 2.3/(p(1-\mu_C))$ layers	Not at bounded gain under strict mixing	Residuals; re-injection in looped models; weight decay and EMA anchors	pending
C2 Alignment (W_V)	Tied value map fails at constrained head dimension	Yes (MLP capacity, width)	W_V; latent KV	pending
C3 Norm and gain control (LN)	Trainability cliffs; logit growth	Partly	Pre-LN, QK-norm, hybrid norm	pending
C4 Write capacity (MLP)	Pure transport cannot write	A write channel restores classification, not retrieval	MLP, W_O	—
C5 Head multiplicity	Loss of subtraction capacity	Yes	Multi-head	pending

The notes on the rows give each row's scope, and most point to a proof that this book gives earlier. Row A1 asks for a way to reduce a head's update in the declared routing class, not for a particular named sink; the gate-times-softmax factorization is that of Sections 8.1 and 8.2, and the budget is that of Chapter 9. The arrow of row A3 is stated per head (Section 10.1). Row A4 rests on the reversible walk that a row-normalized symmetric affinity gives and on the spectrum, in $[0, 1]$, of the kernel produced by tied positive-semidefinite scores (Section 1.2, Proposition 1.1); indefinite scores buy negative eigenvalues. Exact sparsity, row A5, is the subject of Section 6.4. Row B1's margin is the certificate of Section 7.1, in which the logit scale, the radius and the anchor trade against one another (Section 7.5), and row B3 is the two-sided law and its tempering (Sections 9.1 and 9.2). Row C1's price law is that of Sections 2.2 to 2.6 and Section 5.1, and row C4's transport-only statements hold for the pure transport mixer, whose Gram invariant the writing

variants break (Sections 11.1 and 11.7). Row A2's note is taken up in Section 15.5, and row B2's, on bounded state, in Chapter 18.

Three attempts to make the value channel compact or independent did not open room (row C2). In a five-seed repair test on a four-layer transformer trained on dictionary retrieval, a targeted residual beat random directions by only 1.7–1.8× at the middle and last layers (ratios of mean held-out layer-output errors), under a registered bar of 2× in every seed (at the first layer, already near exact, it cleared the bar). A rank-zero arm already solved the task. In a separate execution benchmark with random inputs and newly initialized heads, exactly equivalent untied compact readers (to 1.3×10^{-15}) decoded 1.12–1.81× slower than the dense one, on the hardware it was timed on (Bruce, 2026d).

LayerNorm is a norm-control mechanism, not a certified cap (row C3): its Lipschitz constant is of order $\epsilon^{-1/2}$, where $\epsilon > 0$ is its variance stabilizer (Section 7.5).

14.2 The lattice: a relaxed constraint obliges a protection

Read across its rows, the lattice says that relaxing a constraint of the block buys a capability and obliges a re-supply. Sigmoid attention gains freedom in its row sums and must re-supply a length budget and norm control (row A1); recurrences gain a bounded state and must re-supply recall (row B2); signed mixing gains subtraction and must re-supply a norm cap or per-head normalization (row A2, Section 14.3). Chapter 0 reads the usual answer that the ecosystem co-evolved around the transformer (Section 0.2.1) in these terms: it is not disputed, and each alternative also carries a re-supply obligation for what it relaxes (Table 0.1). Within one class the rows combine into a trichotomy. For normalized nonnegative filters of a fixed mode $0 < \mu < 1$, content survives strengthening if and only if the filter keeps short-path mass on a bounded window (Theorem 2.1). Leaving that class buys content in one of two other currencies: signed coefficients, which leave the stochastic cone (Propositions 1.4 and 2.9 and Section 15.5), or stochastic transport with a stochastic inverse, which exists only for permutations. Each relaxation of the anchor buys one currency and pays its price. We state the trichotomy for this class only, and as a conjecture beyond it.

"No fourth currency" is false in general. The operator $P_L = (1-L^{-2})I+L^{-2}U$, with $U = \mathbf{1}\mathbf{1}^\top/N$, keeps centered gain $(1 - L^{-2})^L \to 1$ at depth L, because it changes with L, outside the fixed-mode premise: it moves toward the identity as depth grows. External gain or a new source of information changes the model as well.

Figure 14.1 draws the lattice. Its first panel is an illustrative construction: relaxing nonnegative weights and unit row sums gives an exact Boolean lattice of four kernels, in which subtraction can raise the row-norm budget and a gate permits abstention, with length calibration where a null competes with the attended keys. Its second panel pairs the six inventory rows with typical mechanisms and with their scope or obligation: an ordinary residual does not guarantee a floor, and normalization alone is not a certificate for the full feedback Jacobian. These are conditional protections, and the low-cost bundle that the figure names is an interpretation under the price model of Section 14.4, not a measured optimum.

14.3 The positivity trade: signed mixing

Nonnegative mixing with unit row sums keeps every output in the hull of the values, since each output is a convex combination of them (Proposition 11.1); a signed kernel gives that guarantee up. Constructions show what each side guarantees and what it forbids (Section 15.5). On random Gaussian score matrices at $N = 24$, softmax keeps every output inside the hull of the values in each

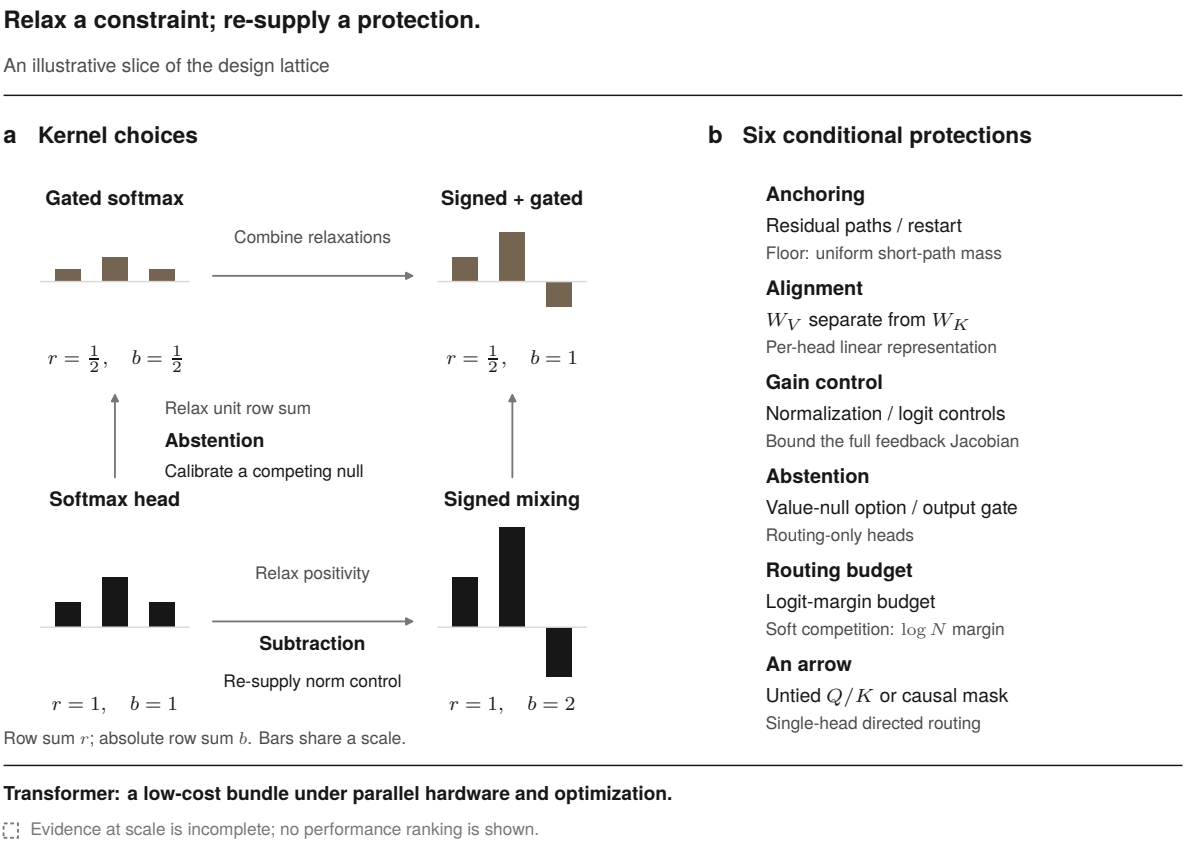


Figure 14.1: Attention designs are organized by the protections they preserve and the capabilities they gain. (a) An illustrative construction shows the exact Boolean lattice obtained by relaxing nonnegative weights and unit row sum. The bottom rows are $(1/4, 1/2, 1/4)$ and $(1/2, 1, -1/2)$; the upper rows multiply them by a gate $g = 1/2$. Bars share a signed scale with zero at the grey line; $r = \sum_j p_j$ and $b = \sum_j |p_j|$. Subtraction can increase the row-norm budget; a gate permits abstention ($g = 0$). Length calibration applies when a null competes with the attended keys. (b) The six inventory rows pair typical mechanisms with their scope or obligation. Short-path mass is the retention condition for normalized nonnegative path laws; an ordinary residual does not guarantee a floor. Normalization alone is not a certificate for the full feedback Jacobian. The arrow statement is per head; compositions can acquire direction. These are conditional protections, not six universally mandatory components. The low-cost transformer bundle is an architectural interpretation under the price model of Section 14.4, not a measured optimum. No empirical measurements are drawn, and evidence at scale remains incomplete.

of five independent seeds, while each of five signed alternatives leaves it, from a softmax with its constant mode removed to the raw, unnormalized scores. Positivity is not a spectral cap, however: softmax's spectral norm exceeds one, 1.048 at the primary seed and 1.048–1.064 over the five, whereas the exact spectral cap, which costs an SVD, sets the norm to one. These are construction tests, not trained-attention results.

The positivity row of the lattice states the trade (row A2 of Table 14.1). Relaxing positivity per head buys subtraction, contrast, high-pass filters and Perron deflation, and it obliges in return a norm cap, that is a spectral contraction, or per-head normalization.

Read through that row, differential attention (Ye et al., 2025), first released in 2024, is signed mixing plus per-head normalization: norm control, not a certified cap. This is one of the readings of the retrodiction table of Section 16.5. Like every reading there, it is an interpretation, not evidence for the lattice, and differential attention is a proposal that practice has not generally adopted.

A frozen construction shows the trade at its simplest. Two strictly positive row-stochastic heads, combined through scalar value and output coefficients $+1$ and -1, mix with signed weights. Their difference annihilates constants, but cancelling constants leaves the remaining gain uncontrolled: the exact spectral cap reduces the contrast gain from 1.6 to 1 (Section 15.5). This is a frozen scalar-head illustration, not a trained-model result or a learning guarantee. A reading of differential attention itself through the trade is left for later work.

14.4 The price list

Section 2.3 priced suppression in collapse: by the price clause of Theorem 2.8, suppressing an interior mode costs collapse, and plain depth pays the least at the suppressions it can attain. For the standard species of Table 1.1 the prices are explicit.

Proposition 14.1 (Prices of the standard species)**.** *Fix a lazy reversible spectrum with content eigenvalue $\mu_C = 1 - \beta \in (0, 1)$ and an interior eigenvalue $0 < \mu_0 < \mu_C$, and put $q = \ln(1/\mu_C)/\ln(1/\mu_0)$ and $r = (1 - \mu_C)/(1 - \mu_0)$. At a suppression $\Sigma = g(1)/g(\mu_0)$ of the mode μ_0, depth pays the collapse ratio $\kappa_{\mathrm{col}} = \Sigma^q$, the residual stream $\kappa_{\mathrm{col}} = \Sigma^r$ and the restart $\kappa_{\mathrm{col}} = 1 + r(\Sigma - 1)$, where $r \geq q$. A constant-share stack pays between depth and the stream, and the restart pays the most.*

The proof is in Section A.15. A machine-precision check reproduces the closed forms to 7×10^{-15} and the ordering at matched suppression, depth no more than the stack, the stack no more than the stream and the stream no more than the restart, on 36 grid points (Appendix B). These are the spectral prices of the price list: what each species of operator pays in collapse to suppress a mode.

The species' prices are one part of a price model. The reading that softmax attention in a pre-normalized block is where every protection is cheap rests on a semi-formal price model: parallel hardware, gradient-based optimization, and prices read from the species (Proposition 14.1) and from the allocator over protections (Sections 16.1 and 16.2) (Bruce, 2026d). Its formal parts price spectral costs and allocations under assumed concave prices; cheapness on the hardware is judged, not computed (Section 0.3.3). This is where the account meets two of the usual answers. Chapter 0 reads the answer that the transformer fit the hardware, large dense matrix multiplications (Section 0.2.1), through the price model: hardware sets the prices, and the conditions are what is being priced. It reads the answer that even attention's data-dependent routing is produced and consumed by matrix multiplications (Section 0.2.2) as part of the price list: other hardware would price other machines differently (Table 0.1).

Chapter 0 states where the price list leads (Section 0.3.3): the transformer is the cheapest point under one price list, not the only one. The machines that pay the same conditions in other currencies

are those of Chapter 19, Bruce (2026e) reads the price list for them, and its full treatment is left for later work.

Underneath the price list sits a family of lower bounds, the conservation of routing information (Proposition 9.12). Whatever the machine, naming one of N stored items costs about $\ln N$ nats, by Fano's inequality (Cover & Thomas, 2006), and machines pay in different currencies: key width, sharpness of the read, state or order. Each bound holds within its own class of machines and prices one currency there, and the bounds set no exchange rate between currencies (Section 9.5).

14.5 The sweet spot is an interpretation

The prices of the alternatives were not measured on a common hardware and optimization budget, so the reading that the standard block is where every protection is cheap is an interpretation under the price model, not a measured optimum (Bruce, 2026d). Nor are the prices independent of the design: hardware, kernels and recipes co-evolved with the transformer, so part of its cheapness is incumbency. The sweet spot is a judgment under one price list, not a computation (Section 0.3.3).

Two measured mechanisms help hold the design where it is, and a third is a hypothesis (Bruce, 2026d). Necessity is real but local, and composition buys it back: compensation (Section 15.1). Protections made into explicit knobs do not beat a block that already pays for them: mechanisms are not knobs (Section 15.4). The third is corners. If prices are concave, optimal allocations sit at corners, since a concave cost is minimized at a vertex (Rockafellar, 1970, §32), and designs would then jump rather than drift; that reading is a hypothesis, because concave pricing of architectures is an assumption, not a measurement (Sections 16.1 to 16.4). Chapter 0 reads the usual answer that what is fixed is a template, not a block (Section 0.2.3), in these terms: what is fixed is the set of conditions and the cheapest current way to pay them (Table 0.1).

The sweet spot, the price list, the retrospective reading of adopted changes (Section 16.5) and the checklist for a successor (Section 19.6) are readings under a model, not statements about networks. They stand or fall with the predictions stated in advance, which Chapter 20 gives in full.

14.6 What holds in trained networks

An interpretation, with nothing to transfer. The sweet spot, the reading that the standard block pays for every condition at once (Sections 14.1 and 14.5): a reading under a semi-formal price model; there is nothing to transfer. The price list (Section 14.4): a reading that joins the hardware account to the conditions; there is nothing to transfer. Relaxing a constraint and re-supplying a protection (Sections 14.2 and 14.3): a reading of the lattice; the individual re-supply obligations are proved in their chapters where stated. That cheapness rests on a price model, not on measured prices (Section 14.5): a statement of scope.

Chapter 15

A flat neighbourhood: compensation and knobs

Chapter 0 states in brief why the neighbourhood of the standard design is flat (Section 0.3.4): two measured mechanisms make local changes rarely pay. This chapter gives both in full. The first is compensation: a necessity theorem binds exactly on its own construction, and a trained network evades it wherever another resource has room (Section 15.1). Section 15.2 says when compensation may be invoked and where it did not apply, and Section 15.3 what it implies for practice; its exception, the anchor, which capacity elsewhere cannot buy back at bounded gain under strict mixing, is stated with its premises in Section 5.5. The second mechanism is that mechanisms are not knobs: a protection made into an explicit knob does not beat a block that already pays for it (Section 15.4). Section 15.5 then takes up a protection the block already has, subtraction across heads, and Section 15.6 the open question of whether compensation grows with scale.

15.1 Compensation

Necessity theorems bound what one constrained map does in one application, and they are exact on their own constructions. Trained systems evade them by composition wherever another resource has room. The measurements behind both sentences are those of Section 6.6: a content map of rank $C-2$ on the proof's own construction, and one of the same rank trained jointly with its readout; a tied value map given spare head dimensions (Section 6.1); MLP capacity against the tax of tying the value map; and heads that buy back subtraction (Section 15.5). The neighbourhood of the standard design is therefore flat. The anchor is the exception, which Section 5.5 states with its premises. Chapter 0 reads the usual answer that the architecture is not the algorithm, that much of what would otherwise need a redesign can be learned inside the weights (Section 0.2.2), as compensation, measured, with an exception the usual answer does not anticipate: the anchor (Table 0.1).

A compensator reconstructs the desired task from the state that a constrained map actually retains (Definition 5.2), and when one exists is itself settled.

Theorem 15.1 (Exact and stable compensability)**.** *For maps $h : \mathcal{X} \to \mathcal{Z}$ and $f : \mathcal{X} \to \mathcal{Y}$, there is a map $c : h(\mathcal{X}) \to \mathcal{Y}$ with $c \circ h = f$ if and only if $h(x) = h(x')$ implies $f(x) = f(x')$, that is, f is constant on the fibres of h. For linear maps $h(x) = Ax$, $f(x) = Bx$ on all of $\mathbb{R}^d$, a linear C with $CA = B$ exists if and only if $\ker A \subseteq \ker B$, equivalently $\operatorname{row} B \subseteq \operatorname{row} A$. On metric spaces there is an L-Lipschitz compensator on $h(\mathcal{X})$ if and only if*

$$d_{\mathcal{Y}}(f(x), f(x')) \le L\, d_{\mathcal{Z}}(h(x), h(x')) \quad \text{for all } x, x'.$$

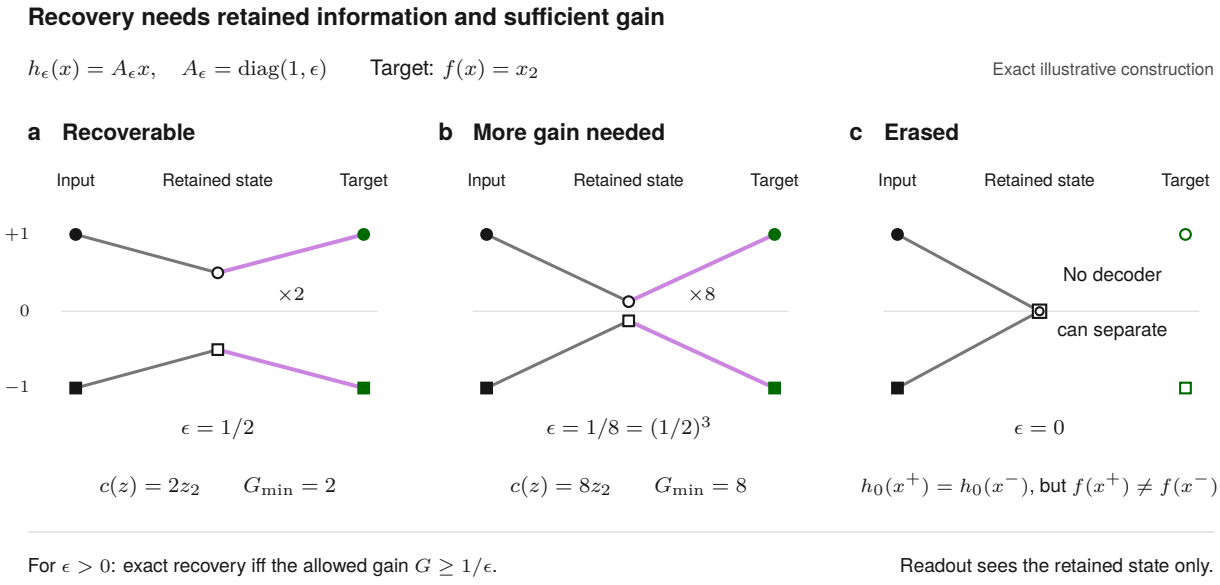


Figure 15.1: **Recovery depends on retained task information and the gain allowed to the compensator.** Exact illustrative construction for $h_\epsilon(x) = A_\epsilon x$, $A_\epsilon = \mathrm{diag}(1, \epsilon)$, and target $f(x) = x_2$ on $\mathbb{R}^2$. Circles and squares track $x^+ = (0, 1)$ and $x^- = (0, -1)$; vertical positions show the second coordinate at each stage on a common scale. (a) At $\epsilon = 1/2$, the readout $c(z) = 2z_2$ restores the target. (b) At $\epsilon = 1/8$, equivalently three applications of $A_{1/2}$, exact recovery requires gain 8. (c) At $\epsilon = 0$, both inputs have the same retained state, so no decoder of that state alone can attain their different targets (open green marks). Orchid lines show the compensating readout; filled green marks are attained targets. For every x, recovery by a scalar Euclidean G-Lipschitz readout is possible exactly when $G \geq 1/\epsilon$ for $\epsilon > 0$. This is not an accuracy measurement or a test of which resources trained networks use.

The proof is in Section A.16. If two inputs with the same retained state have different labels, even an unbounded nonlinear compensator cannot distinguish them without another input. How much gain a compensator needs is a separate matter. Figure 15.1 shows both on an exact construction. The map $h_\epsilon(x) = A_\epsilon x$, with $A_\epsilon = \mathrm{diag}(1, \epsilon)$, keeps the target $f(x) = x_2$ only at the scale ϵ. For every x, a scalar Euclidean G-Lipschitz readout of the retained state recovers it exactly when $G \geq 1/\epsilon$, for $\epsilon > 0$ (Remark 5.7): one halving, $\epsilon = 1/2$, needs gain 2, and three applications of it, $\epsilon = 1/8$, need gain 8. At $\epsilon = 0$ the inputs $(0, 1)$ and $(0, -1)$ have the same retained state but different targets, so the target is not constant on that fibre, and no decoder of that state alone attains both. The construction is an illustration, not an accuracy measurement or a test of which resources trained networks use.

Where the construction binds, the theorems are exact, and the records show it on two constructions (Figure 15.2). On nearest-content retrieval with content dimension $d_b = 16$, a head whose value map is tied to its keys, $W_V = W_K$, recovers retrieval once its head dimension grows past d_b, at 32 and at 128 tokens, with and without value noise; the tied and the separated heads both train their decoder. On a rank construction for $C = 10$ classes, the rank cliff persists when only the readout of the fixed construction is trained. At rank budget 8, a fixed contrast-aligned content map with a fixed diagonal head sits at 90.16%, against the predicted ceiling $(C - 1)/C = 90.0\%$ of its rank-8 construction; the same fixed map with a trained head reaches 90.31%; and a content

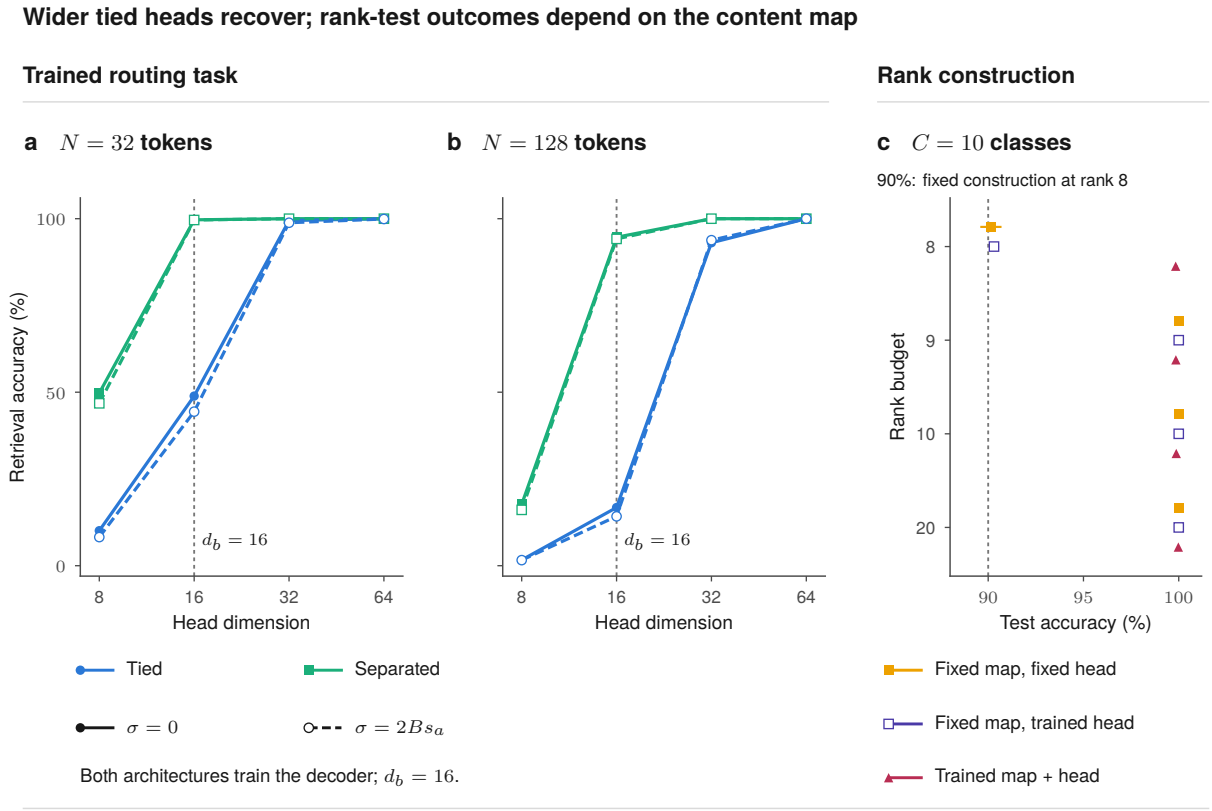


Figure 15.2: **Wider tied heads recover retrieval, while the rank cliff persists when only the readout of the fixed construction is trained.** (a,b) Nearest-content retrieval on the trained routing task at $N = 32$ and 128 tokens, with content dimension $d_b = 16$ and head dimensions $8, 16, 32, 64$ on a logarithmic axis. Both architectures train the decoder; the tied arm shares $W_V = W_K$. Filled solid and open dashed traces show $\sigma = 0$ and $2Bs_a$, respectively, where $s_a = \sqrt{2 \log N}/B$, with spectral budget $B = 1$ on Q/K/V and an unbounded-norm decoder; value noise has per-coordinate standard deviation $\sigma/\sqrt{d_{\text{head}}}$. (c) Test accuracy on the rank construction for $C = 10$ classes: a fixed contrast-aligned map with a fixed diagonal head, the same fixed map with a trained head, and a jointly trained map and head. At rank budget 8 their means are 90.16%, 90.31% and 99.84%. The fixed map's realized ranks are $8, 9, 9, 9$ at the four ordered budgets; the 90% reference applies to its rank-8 construction. All arms use the depth-one pooled deviation with frozen structure map, eight tokens, feature noise 0.4, 2,000 training and 4,000 test examples per seed. Points are means over five seeds; bars are pointwise 95% percentile bootstrap intervals from all 5^5 ordered seed resamples. Intervals can be narrower than markers. All 160 routing-task and 40 rank-construction cells are included; sequence lengths and protocols are not pooled. These retrospective synthetic comparisons do not isolate readout-only rescue in the jointly trained arm.

map and head trained jointly reach 99.84%. Each is a mean over five seeds, and these retrospective synthetic comparisons do not isolate readout-only rescue in the jointly trained arm. The headline numbers of the same records are those of Sections 6.1 and 6.6.

Evasion in trained systems has a currency, and for the value map it has been priced: the MLP ratio at which MLP capacity buys back the tax of tying the value map, in small vision

transformers on CIFAR-100, is given in Section 6.6, as a description at the level of accuracy. Because compensation names its currency, it can be tested, and the book puts a missing signature down to it only under that test (Section 15.2).

15.2 Where compensation did not apply

Compensation does not excuse a missing signature by default. Chapter 0 states the test under which it may (Section 0.3.2): compensation is itself a measured claim with a named currency, such as MLP width, head width or head count, and a model that lacks a protection and shows no signature counts against its row unless removing that currency brings the signature back. That test exists so far for the value map (Sections 6.1 and 6.6). For the anchor, compensation at bounded gain is ruled out under strict mixing (Section 5.5), and the trained test is pending. The deletions to which the principle did not apply, sorted as masked, refuted and void, are left for later work.

15.3 Consequences: architecture search, pruning and LoRA

A flat neighbourhood changes how a proposal should be read. Chapter 0 draws one consequence for architecture work (Section 0.6.1): read small-scale wins with compensation in mind, since if capacity buys back deficits, a gain measured in a small model may shrink in a large one, although that is still a hypothesis (Section 15.6). Head pruning removes a resource outright, and its records so far measure what the removal costs and what fine-tuning the LayerNorm and readout alone buys back. In a pretrained ViT-B, whose top-1 accuracy on 5,000 ImageNet validation images is 81.36%, removing a quarter of each layer's heads leaves 55.52–61.22% under three selection criteria (importance, overlap and random), and removing half leaves 5.46–8.68%. Fine-tuning only the classifier head and the LayerNorm affines of the pruned model, for one epoch on 10,000 images disjoint from those evaluated, brings it back to 65.90–69.44% at a quarter and to 21.54–23.42% at half. These are single-seed records of one model and one currency; the records of the removals alone also cover DeiT-B and a fourth criterion, attention entropy. Low-rank fine-tuning and the other currencies have not yet been measured.
Pending: Which currency repairs a pruned model: recovery after head pruning, under three pruning criteria at a quarter and at half of the heads, when only the LayerNorm and readout, the MLP, attention, low-rank adapters on the MLP or on attention, or the whole model is fine-tuned.

15.4 An explicit knob does not beat the block's own mechanism

The second mechanism concerns the protections themselves. Twelve preregistered interventions each made one protection an explicit, certifiable knob in a standard model: certified fast weights, an anchored mixture-of-experts router, sign-flexible spectral filters, a sink-anchored decoder, margin-instrumented self-distillation, sphere-normalized graph learning, free-value attention, palindromic scheduling, readout-aligned composition, a flow-priced graph transformer, an anchored diffusion sampler and critical-fraction recursive training.

The mechanism layer held in several rows where the theory's hypotheses held, among them certified fast weights, the flat-band filters' price bound, the sink-anchored decoder's envelope and palindromic scheduling, several times to numerical precision, but not in all: it was not met for the flow-priced graph transformer, was inconclusive for self-distillation and the anchored sampler, and its falsifier was realized for critical-fraction recursive training. The capability layer was null or negative in each of the ten capability tests that could be scored. Gains appeared only when

paid for on another axis—the sink-anchored decoder bought length robustness with in-distribution perplexity—or stayed below the registered bar, measured on trained models. Chapter 0 reads the usual answer that the design space around the transformer is flat, since most published modifications did not help when they were re-implemented in one codebase (Section 0.2.3), as explained and measured: by compensation and by this study of knobs (Table 0.1).

Standard components already pay for the protections. The residual is the anchor (Section 5.1), and its ledger in trained pre-normalized stacks is that of Section 5.3; LayerNorm is norm control, although not a certified cap (Section 7.5); and multi-head attention with signed value and output maps subtracts, as Section 15.5 shows. In the router case below, a tuned Switch load-balancing loss already balances its experts.

The study answers a question that theories of deep architectures invite. Such theories name protections, such as an identity path that anchors the constant mode, an independent content map, or a loop gain held below the anchor, and suggest that making them explicit should help. Each of the twelve interventions turned one protection into an explicit, certifiable knob inside a standard model, and each was scored twice against frozen criteria: does the predicted mechanism occur, and does the knob beat a tuned baseline? The design of each row, its protection, knob, mechanism test, capability test and baseline, is left for later work.

The two verdicts can part. Certified fast weights show the mechanism layer at work: they had no divergences in 48 certified runs, where 53 of 273 uncertified runs diverged (Sections 0.3.4 and 18.2), while their capability leg, which lacks the paired runs it needs, is one of the legs still unscored. The row-by-row record, each mechanism test's margin against the capability change relative to the tuned baseline, with the verdict frozen for each row, is left for later work.

One row's records are drawn in full, those of the anchored mixture-of-experts router (Figure 15.3). The anchored router mixes a learned content distribution, with weight c, with a uniform anchor of mass $1 - c$. Its frozen sweeps show a balance–perplexity trade-off, but they do not give a matched router comparison: the anchored router, its greedy-argmax ablation and its content-preservation penalty use GeGLU experts, with 66.9 million parameters, while Switch and no-balance routing use MLP experts, with 48.0 million. Among the MLP-expert runs, Switch with its tuned load-balancing loss keeps the load far more even than routing with no balance: a tuned Switch loss already balances its experts. The frozen verdict on the router remains mixed, and matched expert controls are pending.

Pending: Can a sign-flexible filter shed the anchor? The flat-band case: suppression against constant-mode conditioning for absolutely monotone and sign-flexible graph filters, with the price bound and the region of an anchor-free escape.

Two rows are informative negatives. Hard range separation, which keeps a head's value map orthogonal to the input directions its routing map reads, hurts; and a composition check breaks down in trained product stacks. Their figure, hard against soft range separation graph by graph and the flagged violations of the composition check, is left for later work.

The sink-anchored decoder's trade is measured in Section 8.5: a per-head anchor gate in a 76.7M-parameter decoder trained at 512 tokens cuts perplexity at 2,048 tokens from 88.1 to 59.8 (lower in 2/2 seeds), but raises it at the training length from 25.80 to 26.85 (higher in 2/2 seeds). Length robustness is bought with in-distribution perplexity.

Pending: Whether the sink-anchored decoder's trade survives at larger scale: its gain in perplexity at 2,048 tokens against its cost at the training length, at about 77M and at 350M parameters, beside a tuned length-robust baseline.

Why do explicit knobs not pay? The study's answer is a demarcation: the knobs do not pay because standard blocks already supply the protections, and a tuned baseline already pays for each (Section 0.3.4). Chapter 0 draws the consequence for method (Section 0.6.1): test a protection by

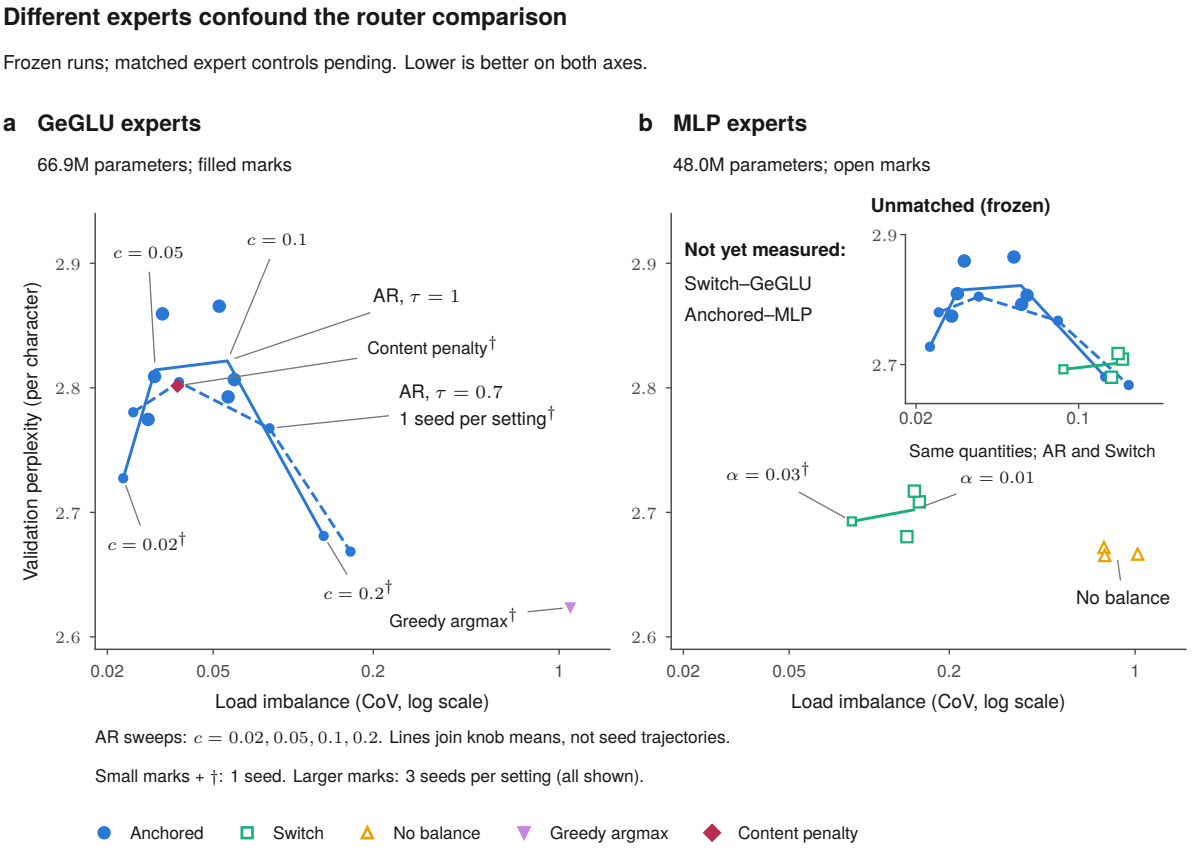


Figure 15.3: **The frozen router sweeps show a balance–perplexity trade-off, but do not give a matched router comparison. (a)** Anchored routing (AR), its greedy-argmax ablation, and its content-preservation penalty use GeGLU experts (66,878,976 parameters; filled marks). **(b)** Switch and no-balance routing use MLP experts (47,955,488 parameters; open marks). For AR, c is the learned-content mixture weight (uniform anchor mass $1 - c$), and τ is the anchored router's softmax temperature; α is the Switch load-balancing auxiliary-loss coefficient. The axes show per-character validation perplexity on WikiText-103 and load coefficient of variation (CoV), averaged over MoE layers and the last 20% of training. All 21 frozen runs are shown. AR at $\tau = 1$, $c \in \{0.05, 0.1\}$, Switch at $\alpha = 0.01$, and no-balance each have seeds 0, 1, 2; the other nine settings have seed 0 only (small marks and †). Lines join arithmetic knob-level means; solid/dashed AR lines denote $\tau = 1/0.7$, with $c = 0.02, 0.05, 0.1, 0.2$ in each sweep. Raw seed spread is shown without inferred intervals. The inset repeats all 16 AR/Switch observations as the original unmatched comparison; it adds no runs. The frozen record gives a mixed result; matched expert controls are pending.

removing it and looking for its signature, not by adding an explicit knob for it. The direct check, one measurement per row of whether the protection is already present in the baseline arm, is left for later work; the anchor ledger of a pre-normalized stack (Section 5.3) and the balance of a tuned Switch router (Figure 15.3) are two such measurements already made.

The study's reach is that of its models. Like the removal signatures and compensation, it comes from controlled runs, mostly below 25 million parameters (Section 0.5). It reports its audit trail beside its results: the artifacts found, the verdicts kept frozen and the legs still unscored. A fuller statement of its limits is left for later work.

15.5 Signed attention is already inside the transformer

One protection the block already has is subtraction. The lattice's positivity row (Section 14.1) says that relaxing positivity per head buys subtraction, contrast, high-pass filters and Perron deflation, and must re-supply a norm cap or per-head normalization (Section 14.3). Its note says why a transformer gains little from the relaxation: differences of heads already realize subtraction inside multi-head attention, so a signed head adds capacity mainly when heads are scarce.

The mechanism is exact for one pair of heads (Figure 15.4). Two strictly positive row-stochastic heads P_1 and P_2, combined through scalar value and output coefficients $+1$ and -1, give $Y = P_1X - P_2X = DX$ for token features X: signed mixing by $D = P_1 - P_2$. In general, two heads suffice for every signed kernel with constant row sums.

Theorem 15.2 (Scalar stochastic heads)**.** *A real token matrix T has a representation $T = \sum_h a_h P_h$ with scalar coefficients and nonnegative stochastic P_h if and only if $T\mathbf{1} = s\mathbf{1}$ for some scalar s. Two heads suffice. The minimum coefficient cost is* $\inf \sum_h |a_h| = \|T\|_\infty$*. For a nonzero matrix with both positive and negative entries, one head cannot suffice.*

The proof is in Section A.16. Matrix rank is therefore not the minimum head count for a scalar signed kernel, and a target whose row sums vary cannot be represented in this class at any head count. The representation treats heads as arbitrary stochastic matrices; low-rank query–key scores impose further restrictions. Cancelling constants leaves the remaining gain uncontrolled. Both D and its exact spectral cap $D_{\text{cap}} = D/\max(1, \|D\|_2)$ annihilate the constant vector $(1, 1)^\top$, but on the contrast vector $(1, -1)^\top$ the gain of D is 1.6, which the cap lowers to 1. Positivity keeps outputs in the hull of the values but does not cap the spectral norm at one, and signed mixing needs norm control, a cap on its norm or per-head normalization (Section 14.3). This is a frozen scalar-head illustration, not a trained-model result or a learning guarantee.

Inside a trained block, heads buy the subtraction back (Figure 15.5). On a controlled task, retrieving the difference $x_a - x_b$ of two tokens' features with one block of width 48 trained for 500 steps, the held-out R^2 of softmax attention rises from 0.46 with one head to 0.998 with four, means over 25 matched runs at each head count. Heads that may instead learn a gate g, the weight of a sign-softmax branch beside softmax, use less signed mixing as heads are added, against a copy-average control, $(x_a + x_b)/2$. Sign-softmax applies softmax to the absolute scores over allowed tokens and restores each score's sign. These are descriptive results for one block and 500 steps, which establish neither convergence nor performance at scale; the gates of individual runs and their trajectories were not retained.

The gates were then tested in two models of six blocks with $H = 4$ or 8 heads per block, a byte-level language model and a CIFAR-10 vision transformer, with every gate started either at $\text{sigmoid}(-2) \simeq 0.119$ or at 0.5 (Figure 15.6). The final gate distributions remain separated by initialization, although separate one-block, one-head difference-retrieval controls open past 0.35

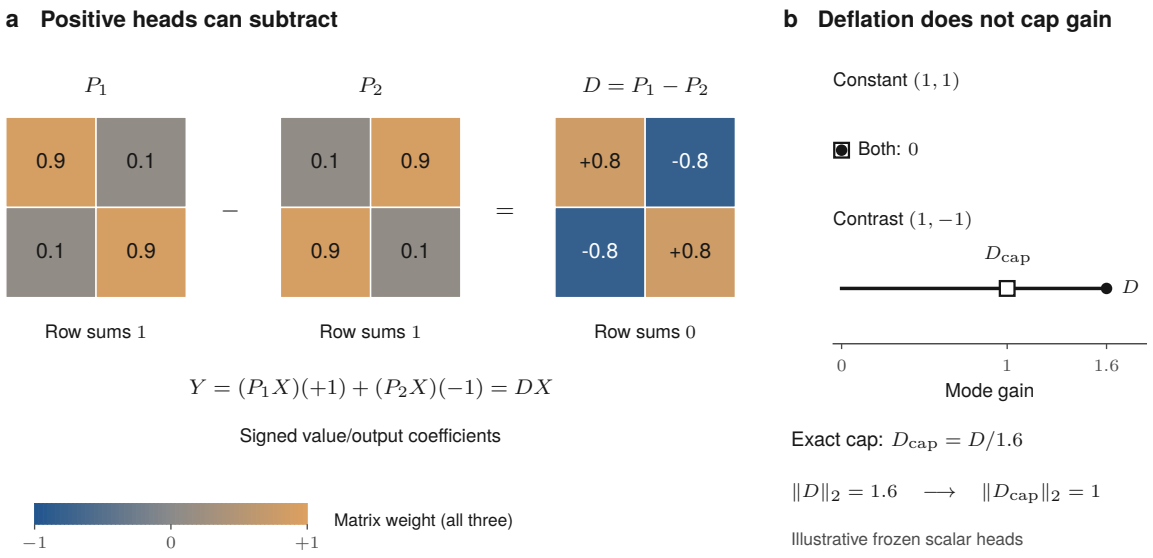


Figure 15.4: **Positive attention heads can implement signed mixing, while cancelling constants leaves the remaining gain uncontrolled.** (a) Exact illustrative construction: the strictly positive row-stochastic heads P_1 and P_2 combine through scalar value/output coefficients $+1$ and -1, giving $Y = P_1X - P_2X = DX$ for token features X. Cell labels and the shared scale show every weight. (b) Filled circles show the gains of D on the constant vector $(1, 1)^\top$ and contrast vector $(1, -1)^\top$. Open squares show the exact spectral cap $D_{\text{cap}} = D/\max(1, \|D\|_2) = D/1.6$: both operators annihilate constants, but the cap reduces contrast gain from 1.6 to 1. This is a frozen scalar-head illustration, not a trained-model result or a learning guarantee.

Table 15.1: **Per-axis paired effects of the positivity relaxation at matched clean accuracy.** The positive arm is a row-stochastic channel mixer inside the crown, the signed arm is signed and approximately capped by power iteration, and the uncapped arm is its signed control. Arm columns are five-seed means; each paired effect has its 95% t interval below it.

Axis	Positive	Signed (approx. cap)	Uncapped	Signed − positive [95% t CI]	d_z	p_{flip}	p_{Holm}	Uncapped − signed [95% t CI]
ECE	0.0528	0.03611	0.0408	−0.01669 [−0.04528, 0.01189]	−0.7252	0.25	1	0.004688 [−0.008616, 0.01799]
OOD corruption drop	0.041	0.0435	0.0387	0.0025 [−0.04127, 0.04627]	0.07092	0.9375	1	−0.0048 [−0.01768, 0.008078]
OOD length ratio	1.317	1.343	1.23	0.02605 [−0.36, 0.4121]	0.08379	0.875	1	−0.113 [−0.6077, 0.3817]
Adversarial correlation norm	1.705	1.735	1.601	0.02995 [−0.4843, 0.5442]	0.07232	0.9375	1	−0.1343 [−0.4222, 0.1536]
Long-horizon drift	1.348	1.465	1.268	0.117 [−0.3031, 0.5371]	0.3458	0.75	1	−0.1968 [−0.6197, 0.2261]
bf16 accuracy drop	−0.0023	−0.0028	−0.0076	−0.0005 [−0.01256, 0.01156]	−0.05147	1	1	−0.0048 [−0.01725, 0.007652]

Positive gaps mean worse. The exact two-sided permutation floor is $2/32 = 0.0625$. No realised operator norms were recorded. No axis qualifies as a robustness cost or an approximate-cap benefit under the paired t interval together with a Holm-adjusted $p < 0.05$. The sensitivity analysis excludes seed 4, whose clean gap is $0.0615 > 0.05$, and finds no cost axis.

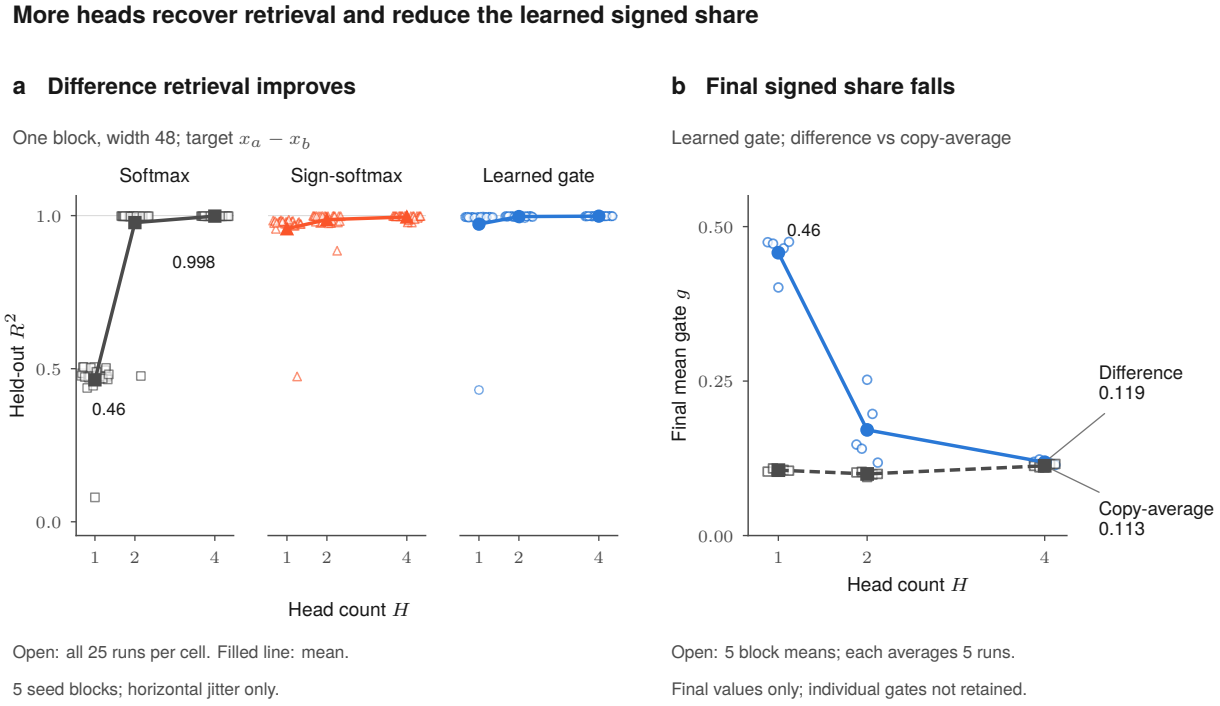


Figure 15.5: **Softmax difference retrieval improves with head count, while learned gates use less signed mixing in this controlled task.** (a) Held-out R^2 for $x_a - x_b$ at $H = 1, 2, 4$ heads. Open marks show all 25 matched runs per cell (five seed blocks of five runs; 512 evaluation examples per run). (b) Final gate g, the weight of the sign-softmax branch in each head: difference retrieval uses circles and solid lines; the copy-average control, $(x_a + x_b)/2$, uses squares and dashed lines. Open marks are the five retained block averages, each averaging five runs and their heads. In both panels, filled lines are arithmetic means and jitter is horizontal. All models have one block, width 48 and 500 training steps; gates start at sigmoid$(-2) \simeq 0.119$. Individual-run gates and trajectories were not retained. These descriptive results do not establish convergence or performance at scale.

in all five seeds after 500 steps. None of the 20 paired model starts meets the preregistered 0.2 mean-logit agreement tolerance, so these endpoints do not establish gate convergence; trajectories were not retained. Averaged over its heads, the final gate lies below one half, on the softmax side, in every one of these runs from either start, between 0.03 and 0.29 over the 40 runs, though the values depend on where the gates began. In the language model the softmax baselines meet their recorded task threshold in 0/5 and 1/5 runs at $H = 4$ and 8, against 5/5 at each vision head count.

Under a norm cap, signed mixing stays bounded at depth in frozen constructions (Figure 15.7). Capped signed operators, with spectral norm at most 0.95 by construction, contract through depth 48, both when every layer draws a fresh operator and when one operator is repeated at every layer, while the uncapped signed control grows; a doubly-stochastic positive reference is shown beside them, and the comparison is repeated at depth 12 for eight times the length. Each value is the recorded median of seven construction draws from one retained seed. These are mixing-only probes without LayerNorm or learned weights: boundedness and decay do not establish that task information is retained, and the capped operators are not evidence that numerical-radius bounds compose across distinct layers.

In isolation, signed mixing does buy expressivity (Figure 15.8). On Texas, Wisconsin and a

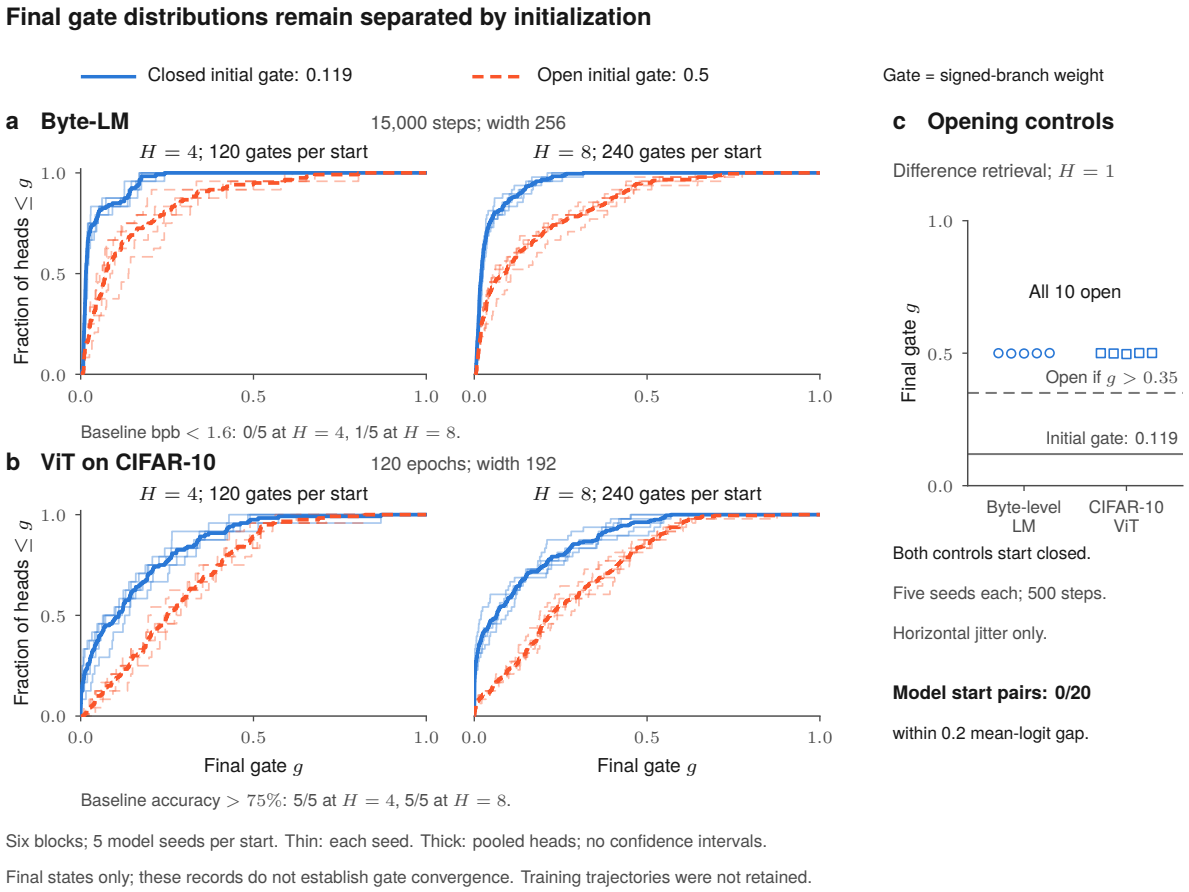


Figure 15.6: **Final gate distributions remain separated by initialization despite responsive opening controls.** (a,b) Empirical cumulative distributions of the final per-head gate g, the coefficient of sign-softmax in $(1-g)\,\mathrm{softmax} + g\,\mathrm{sign\text{-}softmax}$, for the byte-LM and CIFAR-10 ViT. Blue solid and orange dashed curves start at $g = \mathrm{sigmoid}(-2) \simeq 0.119$ and $g = 0.5$, respectively. Thin curves show each of five model seeds; thick curves pool their heads. Each model has six blocks and $H = 4$ or 8 heads per block, giving 120 or 240 gates per initialization, respectively. Every retained per-head value is used at its recorded four-decimal precision; heads are not independent replications and no confidence intervals are estimated. (c) All five seeds of each separate one-block, one-head difference-retrieval control open past 0.35 after 500 steps; marks have horizontal jitter, with initial value and opening threshold shown as solid and dashed lines. Gate optimizer groups have zero weight decay and learning-rate settings 0.01 (byte-LM) or 0.02 (CIFAR-10 ViT). The LM/ViT runs in (a,b) use cosine schedules; the controls in (c) use constant-rate AdamW. None of the 20 paired model starts meets the preregistered 0.2 mean-logit agreement tolerance. Softmax baselines meet their recorded task threshold in 0/5 and 1/5 LM runs at $H = 4$ and 8, versus 5/5 at each ViT head count. These endpoints do not establish gate convergence; trajectories were not retained.

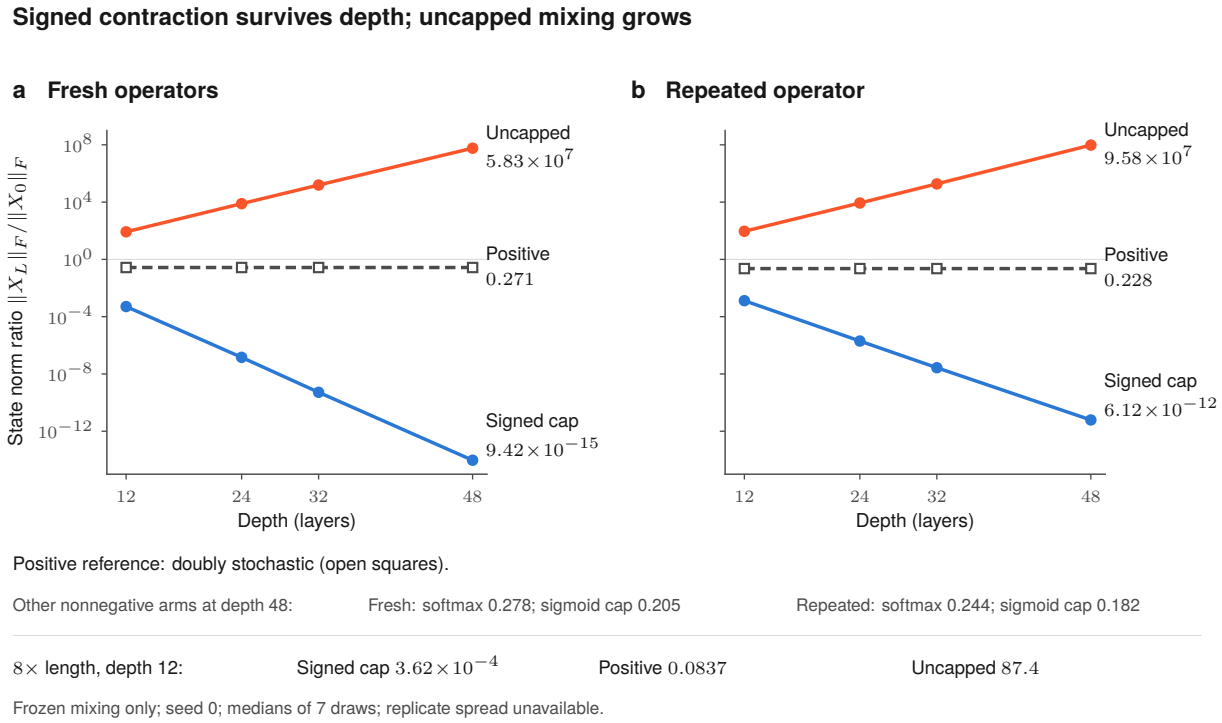


Figure 15.7: **For these frozen mixing constructions, capped signed operators contract through depth 48 while the uncapped signed control grows.** Panels (a) and (b) show the Frobenius state-norm ratio $\|X_L\|_F/\|X_0\|_F$ on a shared logarithmic scale for independent fresh operators and for one operator repeated at every layer, respectively ($N = 16$, eight features). Filled circles show the signed capped arm and the uncapped control; open squares show the doubly-stochastic positive reference. The two other nonnegative arms' depth-48 values are printed below. The bottom row gives depth-12 results from a separate run at $N = 128$ (8× length). Each value is the recorded median of seven construction draws from the sole retained seed, seed 0; replicate values and uncertainty intervals were not retained. The signed capped arm has spectral norm at most 0.95 by construction; this is not evidence that numerical-radius bounds compose across distinct layers. These are mixing-only probes without LayerNorm or learned weights: boundedness and decay do not establish task-information retention.

synthetic heterophily graph, a signed filter keeps a test-accuracy advantage over a matched positive low-pass filter at high polynomial order, as ten-seed means with their retained 95% paired-bootstrap intervals; the order K is the maximum polynomial degree, not a count of attention blocks. On difference retrieval, with single-head blocks of width 48 trained for 500 steps and with residual, normalization and MLP paths, softmax catches up with depth and solves the task at depth 3, while the ℓ_1-signed arm is variable. The shallow advantage is therefore not an impossibility result for the full block.

What was positivity protecting? No robustness cost of the approximately capped signed mixer is detectable with five seeds on the crown substrate. In a post-hoc paired analysis across all six axes (Table 15.1), no paired t 95% interval lies wholly above zero together with a Holm-adjusted one-sided $p < 0.05$. The same criterion finds no demonstrated benefit of its approximate cap over the uncapped control. This does not establish equivalence: the positive and approximately capped

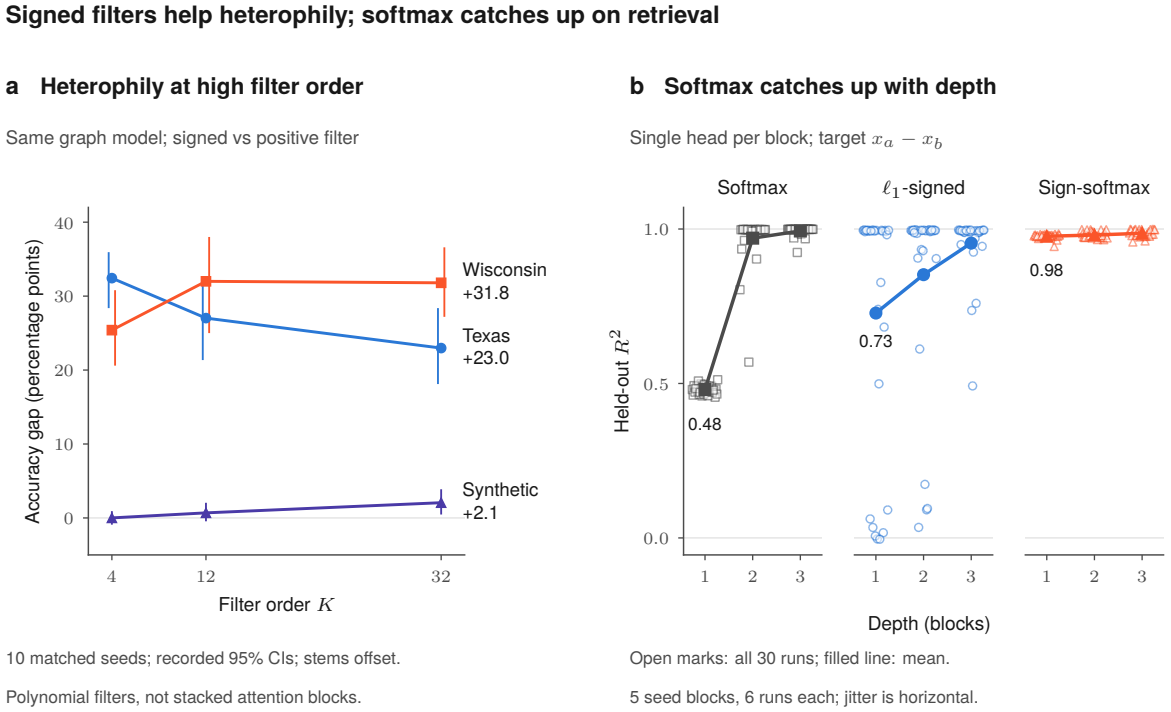


Figure 15.8: **Signed filters retain a heterophily advantage at high order, while softmax catches up on difference retrieval with depth.** (a) Test-accuracy gaps of a signed filter over a matched positive low-pass filter on Texas, Wisconsin and a synthetic heterophily graph. Symbols denote ten-seed means; whiskers are the retained 95% paired-bootstrap intervals (5,000 resamples). Texas/Wisconsin stems are shifted left/right by 1.2 pt for visibility; means keep their true K. All nine dataset/order cells from the sole retained outer seed, seed 0, are shown; K is polynomial order, not a count of attention blocks. (b) Held-out R^2 for difference retrieval, $x_a - x_b$, using single-head blocks with residual, normalization and MLP paths. Open marks show every run, with horizontal jitter; filled marks and connecting lines show descriptive means over five outer seed blocks of six matched runs each, evaluated on 512 examples per run. All three normalizations and depths are shown on the same scale; labels give depth-1 means. The ℓ_1-signed arm is variable, and softmax solves this task at depth 3: the shallow advantage is not an impossibility result for the full block.

signed crown arms' clean accuracies are 0.31–0.38, below the pre-LN softmax reference's 0.39–0.42. Excluding the least-matched seed does not reveal a robustness cost. The construction-level cost that remains is the capped signed arm's mean hull-overshoot score (0.30 versus zero for softmax). The score adds the largest excursions above the input maximum and below the input minimum, then divides by the input range. It is averaged over ten length-16 constructions, with three scalar probes each, from one retained outer seed. This is separate from trained robustness; positivity keeps every output in the hull of the values (Section 14.3).

Pending: Head count at scale: the perplexity gain of signed mixing over softmax, and the learned signed share, against head count at 125M and 350M parameters.

15.6 Does compensation grow with scale?

Compensation predicts that many deficits shrink with capacity; for a deficit that does, its price measured at small scale is an upper bound on what it costs a large model, not an estimate of it (Bruce, 2026d). This is a hypothesis about trained networks at scale, and its registered test is pending. If compensation does grow with capacity, it would also explain why the small-scale gains of alternatives fade with scale (Section 0.3.4). Chapter 0 reads two usual answers through it (Table 0.1). That scale sorts architectures, so far in the transformer's favour (Section 0.2.3), gains a mechanism, as a hypothesis. That proving an alternative at scale costs too much (Section 0.2.1) is refined: small-scale wins can fail validity checks (Section 18.5), and compensation predicts that deficits cost less at scale, which is pending.

Pending: The registered test of compensation at scale: the tax of tying the value map against MLP ratio, width and depth, in vision transformers on ImageNet-100 and in a 125M-parameter language model, against the budget measured at small scale.

15.7 What holds in trained networks

Compensation carries its label once, where it is first measured (Section 6.7): it is measured in trained networks. Its exception, the anchor, carries its own label in Section 5.7.

Exact for every network of the stated class. Signed mixing from positive heads (Section 15.5). A representation identity: two stochastic heads give any signed kernel with constant row sums.

Measured in trained networks. Explicit knobs against a tuned block (Section 15.4). Measured on twelve preregistered interventions in standard trained models.

Pending: there is nothing yet to transfer. Whether deficits cost less as models grow (Section 15.6). A hypothesis about trained networks at scale; its test is pending.

Chapter 16

Corners, rediscovery and the blueprint

Chapter 0 asks two questions of the design's recent history in brief (Sections 0.3.5 and 0.3.6): do designs jump rather than drift, and how do the changes that practice adopted read in retrospect? This chapter develops both. Sections 16.1 to 16.4 take the first, which is a hypothesis: the prices of protections, the theorem that under concave prices the best allocation is a corner of the feasible set, what the recorded designs can and cannot establish, and the controlled comparison of jump against drift that has still to run. Section 16.5 reads the adopted changes retrospectively, as moves in the lattice of Chapter 14. Sections 16.6 and 16.7 run the inventory forward, assemble the smallest block that meets it, and give the evidence recorded for that block and the tests still pending.

16.1 Pricing protections

The corner hypothesis asks how a design budget is best spread across protections, so it needs their prices. They are the prices of the amplifier species (Section 14.4): at a matched suppression of an interior mode, depth pays the least collapse and the restart the most, with the residual stream between.

Write the collapse ratios of Table 1.1 at a common strength t: the depth L, the stream's total update σ, or $t = \alpha/(1-\alpha)$ for a restart with continuation probability α; $\beta = 1 - \mu_C$ is the content gap. The anchored species' premium is robustness at a fixed strength t. The restart's log-price sensitivity to the content gap is $t/(1+t\beta) \le 1/\beta$, against t for the residual stream and $t/(1-\beta)$ for depth, so the ratio of stream to restart is exactly $1 + t\beta$. The clause "at fixed strength" matters. At matched suppression $1/x$ of a nuisance gap $\beta_n > \beta$, the gap of the mode to be suppressed, the same ratio is

$$\frac{\log x}{x-1}\Big(1 + \frac{\beta}{\beta_n}(x-1)\Big),$$

which is below one whenever $\beta/\beta_n < 1/\log x - 1/(x-1)$; it is 0.716 at $x = 10$ and $\beta/\beta_n = 0.2$, by direct differentiation.

These prices, and the premium, are the results on pricing that the corner hypothesis builds on, and a synthetic check verifies them. Over 174 settings of the two gaps and of the suppression, the closed forms agree with the gain calculations to a maximum relative error of 1.8×10^{-15}, and the prices keep their order at matched suppression, depth no more than the stream and the stream no more than the restart, with no inversion. The log-price sensitivities match finite differences to a maximum relative error of 1.2×10^{-8}, and at a common strength the ratio of stream to restart equals $1 + t\beta$ to a maximum relative error of 3.3×10^{-9}. When the performance curves are uniformly strongly concave over the strengths compared, and their slopes have bounded sensitivity to

conditioning, the control result gives a sufficient condition for a dial that uses measured conditioning to beat every fixed strength on average: the best strengths must vary enough across settings relative to the gauge error. Three tests in graph label propagation and synthetic node classification form a ladder: a dial fed a regressed estimate of the best strength, one fed a measured estimate, and a certified allocator. The condition held in none of the three, as recorded. The regressed dial lost to the best fixed strength, the measured one fell short of it within the tolerance set in advance, and the certified allocator beat the best fixed configuration although the condition did not hold, which a condition that is sufficient only allows (Section 16.6). A fourth rung, an optimizer gated between two update rules, had the best mean over three seeds after 3,000 updates on all three tasks: small vision and character-level language models, and synthetic deep linear regression. Its record states no crossover, and only the ordering of its policies is interpretable.

16.2 From concave prices to corner designs

If prices are concave, allocations of protection strength sit at corners. The corner hypothesis rests on this exact result about allocations, which holds on the declared feasible set.

Proposition 16.1 (Corners under concave prices)**.** *Let the strengths $t = (t_1, \dots, t_A)$ of A protections satisfy caps $0 \le t_a \le u_a$, where a cap may be infinite, and one weighted budget $\sum_a \beta_{n,a} t_a = T$ with protection rates $\beta_{n,a} > 0$, and let this feasible set be nonempty. If the price $C(t)$ is concave on it, C attains its minimum at a vertex, and every vertex has at most one coordinate strictly between its bounds. If C is strictly concave, every minimizer is a vertex.*

The proof is in Section A.17. That a concave function on a polytope attains its minimum at a vertex is classical (Rockafellar, 1970, §32). The proposition does not say which vertex is best, and greedy ordering can fail to find it under capped anchored pricing, the restart's log-price $\log(1 + \beta_{C,a} t_a)$ on each axis a, so the allocator enumerates corners.

Figure 16.1 shows where greedy ordering fails. On 24 capped anchored-price instances with two to four axes, it chooses a suboptimal corner in five, by 0.0017 to 0.3892 log-price units, and no saved feasible grid point is cheaper than the best corner. In all 36 constructed instances, capped or not, an interior split never beat the corner. This finite-grid check is distinct from the vertex theorem, and its prices are filter log-prices, not measurements of trained architectures: it is a synthetic check.

Concave pricing admits a vertex optimum, but it does not select a preferred fractional allocation (Figure 16.2). On a capped two-axis budget, a strictly concave cost is minimized at the vertices, and as a linear price tilt θ crosses zero its optimizer switches from one vertex to the other: with the feasible set fixed, the best design moves only where the costs of the two vertices cross. A strictly convex cost has an interior optimum that moves continuously instead. Concavity guarantees a vertex minimizer, and strict concavity is needed to exclude all others. These are exact illustrations, not measurements; the premise that architectures are priced this way remains to be tested.

Nor does concavity select a fraction on a single design axis. On an attention-fraction axis, concave pricing predicts 0, 1 or a binding cap, not a particular fraction such as 1/4, and the level set $\{1/2, 1/3, 1/4, 1/6, 1/8\}$ used to score future releases (Section 20.3) is a convention, not a consequence.

To say anything about architectures, the theorem needs a bridge from allocations of protection strength to architecture axes. The corner reading for architectures rests on concave pricing of protections at the level of architectures, which Chapter 0 states as an assumption, not a measurement (Section 0.3.5); the bridge itself is left for later work.

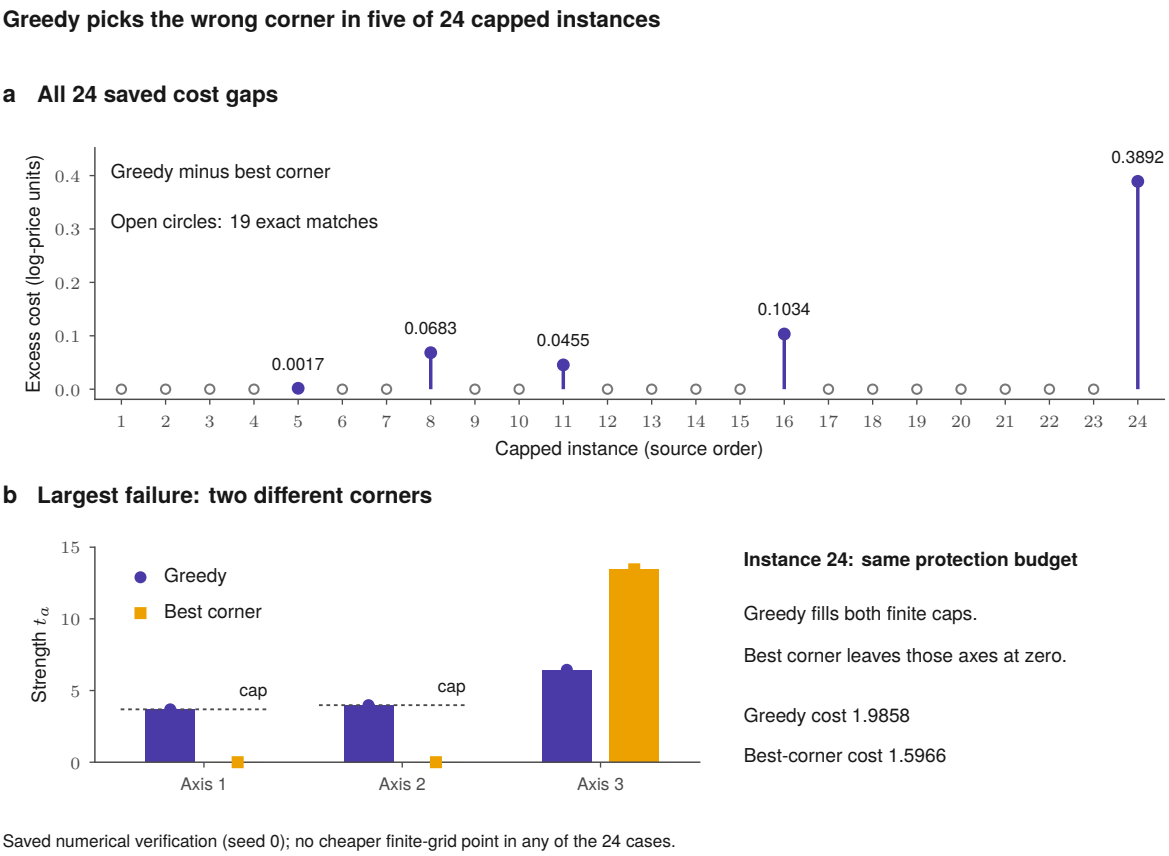


Figure 16.1: **Greedy ordering chooses a suboptimal corner in five of 24 capped anchored-price instances.** (a) Saved greedy cost minus the minimum over enumerated corners, in source order; open circles mark the 19 exact matches, and labels give the five positive gaps (0.0017–0.3892 log-price units). (b) Instance 24, the largest recorded gap: greedy fills the two finite caps, while the best corner puts all protection on axis 3. Bars and circle/square endpoints show the saved strengths t_a, including zeros; dashed lines mark caps. Both allocations meet the same weighted budget $\sum_a \beta_{n,a} t_a = T$, with costs $C(t) = \sum_a \log(1 + \beta_{C,a} t_a)$ of 1.9858 and 1.5966. Here $\beta_{n,a}$ and $\beta_{C,a}$ are protection and price rates. These are 24 numerical verification instances with 2–4 axes, seed 0; exact per-instance costs have no sampling interval. No saved feasible grid point is cheaper than the best corner (60, 36 or 20 points per axis for 2, 3 or 4 axes). This finite-grid check is distinct from the strict-concavity vertex theorem (Proposition 16.1): at most one coordinate is between its bounds. These are filter log-prices, not trained-architecture measurements.

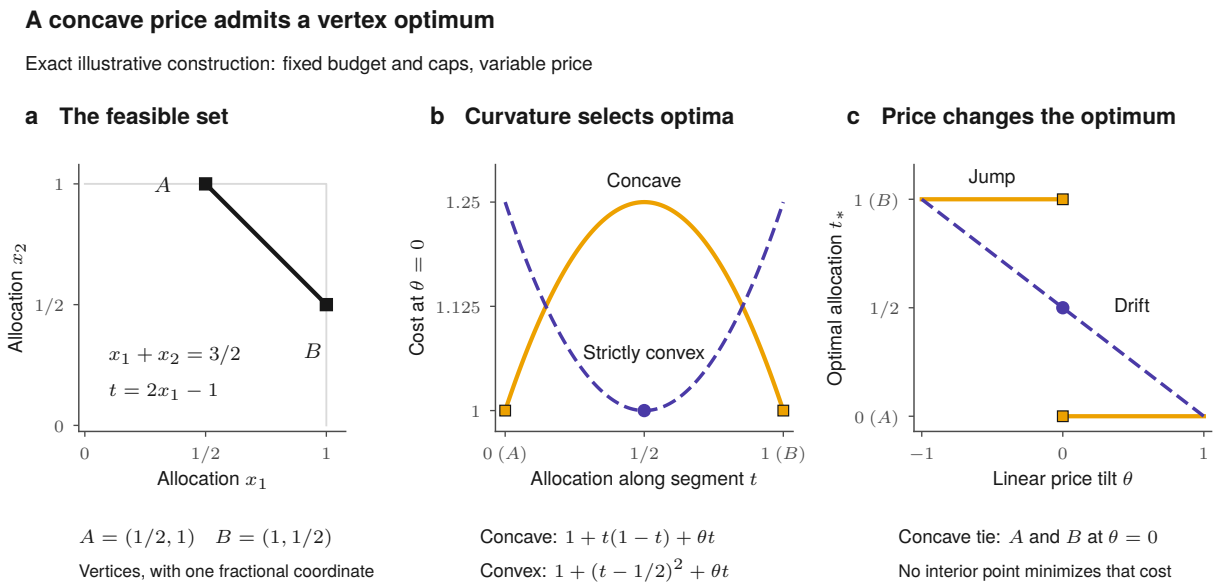


Figure 16.2: **Concave pricing admits a vertex optimum, but does not select a preferred fractional allocation.** (a) Exact illustrative construction on the capped budget set $0 \le x_1, x_2 \le 1$, $x_1 + x_2 = 3/2$. Its vertices $A = (1/2, 1)$ and $B = (1, 1/2)$ each have one coordinate between its bounds; $t = 2x_1 - 1$ parameterizes their segment. (b) At $\theta = 0$, the strictly concave cost $1 + t(1-t) + \theta t$ has its minima at the vertices (solid amber, squares), while the strictly convex cost $1 + (t - 1/2)^2 + \theta t$ has an interior minimum (dashed violet, circle). (c) With the same fixed feasible set, the concave optimizer switches from B to A as θ crosses zero; both vertices, and no interior point, minimize at the crossing. The convex optimizer $t_* = (1 - \theta)/2$ varies continuously for $-1 \le \theta \le 1$. In general, concavity guarantees a vertex minimizer; strict concavity is needed to exclude all nonvertex minimizers. All geometry, costs and optima here are exact illustrations, not measurements of trained architectures. The architecture-pricing premise and deployed behavior remain to be tested.

16.3 What the recorded designs cannot establish

Deployed hybrids sit at or next to small-denominator attention fractions. Qwen3-Next keeps full attention in 12 of its 48 layers, exactly 1/4, and Kimi Linear in 7 of 27, a near miss: its nominal 3:1 ratio of linear-attention to full-attention layers is realized as 7 full-attention layers of 27 (Qwen Team, 2025; Kimi Team, 2025). The record is read from the models' published sources. It is descriptive, and it was compiled by those holding the hypothesis (Bruce, 2026d).

Two records score this saturation, the signature of allocations that sit at levels rather than between them (Bruce, 2026d). The earlier is a curated table of 13 rows that carries review flags and rounds Kimi Linear's fraction to 1/4; the later re-records six systems from their published sources and keeps 7/27. Both score the values by their mean distance to the nearest level, normalized by the spacing of the levels there, against the continuous uniform null used when they were recorded, which places each value uniformly at random within its axis's range. The levels are fractions $1/k$, dyadic loop counts and canonical depths. The six systems are the two hybrids above, a looped language

model (Zhu et al., 2025), and ViT-Base, ViT-Large and ViT-Huge (Dosovitskiy et al., 2021). Their mean distance is 0.03, against 0.32 for the null. For the looped model and the vision transformers, the level each is compared with is its own value: the trained loop cap and the canonical depths. For the two attention fractions alone the mean distance is 0.09, against 0.33 for the null. Neither record scores the designs against nulls that respect integer layer counts and block periods.

What the record cannot establish is the hypothesis. It is descriptive and small, and it was compiled by those holding the hypothesis. Its null is coarse: a uniform draw is a rough model of optimization in the interior, and the discreteness of design space, integer layer counts and block periods among it, inflates the apparent clustering at the levels, which is why these records are read descriptively. The theorem's premise, concave pricing at the level of architectures, is an assumption, not a measurement (Section 16.2). On a single attention-fraction axis it predicts 0, 1 or a binding cap, never a fraction such as 1/4 in itself, so a small fraction counts as a corner only by the scoring convention. The deployed fractions therefore do not yet discriminate between corners and drift. A test needs measured prices, or the controlled comparison of jump against drift (Section 16.4).

A larger record, with each design verified against its source, would give for each system its release date, its layers, the indices of its attention layers or its split of heads, its block period and its loop cap. Attention fractions within a family of hybrids could then be read against release date, as plateaus and jumps between levels $1/k$, or as drift. That record is left for later work.

16.4 Jump or drift?

Pending: The controlled comparison of jump against drift: a learned attention share against its price, under linear and under convex pricing, with a control whose optimum is planted in the interior.
Pending: The corner reading of the qualified hybrid sweep of Section 18.4: quality, robustness at length and state size against the attention fraction, and the fraction that optimizes quality at the training length against the price of state.

16.5 Rediscovery

In retrospect, the changes to the mixing layer and its block that practice has proposed or adopted over the last decade read as moves in the lattice of Chapter 14 (Bruce, 2026d). Table 16.1 gives fourteen groups, dated by first public release: the anchor and a choice of mixing budget; gain control; abstention, patched inside unit row sums, given a null with a fixed budget, or made query-dependent; row sums dropped, with the budget and norm re-supplied; the routing budget; signed mixing with per-head normalization; bounded state, only with contracting writes; a recall reserve; a compact content channel; density relaxed with learned support; the anchor on the recursion axis; and its absence in looped models without re-injection. Several rows are proposals that practice has not generally adopted, the hybrids' corner is the hypothesis of Sections 16.2 to 16.4, and every reading is an interpretation, not evidence for the lattice. Chapter 0 reads two of the usual answers through the table (Table 0.1). That the transformer absorbs its challengers as components (Section 0.2.3) it explains: most of what gets absorbed into the mixing layer re-supplies a named protection, while mixtures of experts, a change to the MLP, lie outside the lattice. That history put the right objective, data and hardware in one place (Section 0.2.1) it refines: the path was historical, but its endpoint, read this way, looks constrained. Both are readings.

Figure 16.3 dates 33 sources of the first 13 groups, without Yang et al. (2024b), by their first public year, that of an arXiv v1, a blog or a model-card release, not the year of adoption. The

Table 16.1: Retrodiction: changes to the mixing layer and its block, proposed or adopted, read as lattice moves. Years are those of the cited sources' first public release, not of adoption, and are not a claim of priority; several rows (sigmoid and differential attention, looped models) are proposals that practice has not generally adopted. The last column names the rows of the lattice (Tables 14.1 to 14.3). Readings are interpretations, not evidence for the lattice.

Practice change (first public)	Lattice reading	Row
Residual connections (2015) (He et al., 2016); Pre-LN (2019–2020) (Wang et al., 2019; Xiong et al., 2020)	The anchor, and a choice of mixing budget: linear for Post-LN, logarithmic (critical) for Pre-LN	C1
QK-norm (2020), z-loss (2022), σReparam (2023), logit soft-capping (2024) (Henry et al., 2020; Chowdhery et al., 2023; Zhai et al., 2023; Gemma Team, 2024)	Gain control: bound s in $\kappa_{\mathrm{f}} \leq 2sR^2$ (z-loss acts on the output logits, a block-level reading)	B1, C3
Attention sinks and registers (2023) (Xiao et al., 2024; Darcet et al., 2024)	Abstention patched inside unit row sums	A1
Off-by-one normalization (2023); learned sink logits (2025) (Miller, 2023; OpenAI, 2025)	A null with a fixed budget	A1, B3
Output-gated attention (2025) (Qiu et al., 2025)	Query-dependent abstention	A1
Sigmoid attention with $-\log n$ and hybrid norm (2024) (Ramapuram et al., 2025)	Row sums dropped; budget and norm re-supplied	A1, B3, C3
$\log n$ scaling (2021), YaRN temperature (2023), scalable softmax (2025) (Su, 2021; Chiang & Cholak, 2022; Peng et al., 2024; Nakanishi, 2025)	Routing budget beyond the trained length, paid with positions kept in distribution (Section 17.4)	B3
Differential attention (2024) (Ye et al., 2025)	Signed mixing plus per-head normalization (norm control, not a certified cap)	A2
Linear attention (2020), state-space models (2021–2023), (gated) delta rules (2021–2024) (Katharopoulos et al., 2020; Gu et al., 2022; Gu & Dao, 2024; Schlag et al., 2021; Yang et al., 2024c, 2025)	Bounded state only with contracting writes ($a < 1$)	B2
Hybrids at or near $1/k$ attention (2024–2025) (Lieber et al., 2024; MiniMax, 2025; Qwen Team, 2025; Kimi Team, 2025)	Recall reserve; a corner if prices are concave	B2; corners
Latent KV (2024) (DeepSeek-AI, 2024)	Compact content channel	C2
Sparse attention with learned blocks (2025) (Yuan et al., 2025; Lu et al., 2025)	Density relaxed with learned support	A5
Looped models with input re-injection (2022–2025) (Bansal et al., 2022; Yang et al., 2024b; Geiping et al., 2025)	Anchor on the recursion axis	C1
Looped models without re-injection (2025) (Zhu et al., 2025)	No anchor on the recursion axis; accuracy is highest around the trained loop count	C1

corpus is not exhaustive, the readings are interpretations, not evidence for the lattice, and listing a method does not certify that its writes contract.

The first row has an exact counterpart in the mean-field model at initialization, which Section 2.6 proves. Post-LN resets the stream after each layer, so its mixing budget grows linearly with depth and forgetting is exponential; Pre-LN's budget grows logarithmically, which is critical; depth-scaled branches, as in GPT-2's initialization and depth-μP, keep it bounded; and randomly initialized transformers with bidirectional attention match the predicted exponent at width 512, averaged over three seeds, with a largest deviation of 0.041. The Post-LN cliff of Section 5.2 remains a trainability observation.

The readings are retrodictions. Each names a lattice quantity, but none was predicted before the change was adopted, so none counts as evidence for the lattice (Bruce, 2026d). The predictions that can are those stated in advance, which Chapter 20 gives.

The lattice is silent on much of what practice adopted. Changes to the MLP (gated linear units, mixtures of experts), to the implementation (fused attention kernels, quantized caches) and to the block's topology (parallel attention and MLP branches) do not appear in it, because they neither remove nor re-supply a protection of the mixing layer. Grouped-query attention reads, like latent KV, as a cheaper content channel. Rotary positions enter through row A3 of Table 14.1, where untied queries and keys let them circulate, and through row B3, the routing budget, which they pay in trained causal language models inside the trained range and, once interpolated, beyond it. We found no adopted change to the mixing layer that removes a protection without re-supplying it, but this search was ours, not blind, and it is a retrodiction too (Bruce, 2026d).

16.6 From map to blueprint

Run forward, the inventory is a constraint set, and the smallest block that satisfies it can be assembled and tested (Bruce, 2026d). The inventory asks for an independent content map (Section 6.1), short-path mass (Chapter 5), a $\log N$ routing budget (Chapter 9), an abstention channel (Chapter 8), gain control (Chapter 7) and, outside the closed-form inventory, interleaved nonlinear writing (Remark 11.2). The minimal certified block uses one symmetric structure map, a $1/L$-scaled residual update with an independent value map of rank at least $C-1$, for C classes, and appended register tokens; its decoder form uses separate queries and keys and the causal mask, which supply the arrow (Chapter 10), and a $\log N$ logit scale. In trained 125M-parameter causal language models such a scale helped beyond the training length on interpolated positions, and raised long-context perplexity on top of YaRN's own temperature, as did YaRN's temperature alone without interpolation (Section 17.4).

Symmetric scores make every layer reversible by construction (Section 1.2), and detailed balance, measured on all 50 trained minimal encoders, holds to 10^{-12} on every layer. Completed with interleaved MLPs and Pre-LN placement and matched to ViT-Tiny's parameter count within 0.04%, the block beat ViT-Tiny after 100 epochs in 20 of 20 paired runs on CIFAR-10 and CIFAR-100, by 1.8 to 3.9 points averaged over the five seeds of each setting (Bruce, 2026d). This is tiny-scale evidence, not a scaling claim; the comparison at 22 million parameters waits for tuned baselines (Section 16.7).

Assembling the minimal object that satisfies the inventory gives, within four corrections, the Pre-LN transformer: explicit registers where practice lets sinks emerge, one symmetric structure map where directed routing is not needed, a $\log N$ budget where practice retrofits one, and a residual update scaled by $1/L$ where Pre-LN is only critical (Bruce, 2026d). We offer this as a reading of an assembled constraint set. It is not a derivation of the transformer, and the inventory it assembles is conditional. Since the inventory was assembled by asking what each part of the block protects



Figure 16.3: **Selected proposed or adopted changes map to recurring protections in a retrospective reading.** Each dot gives the first-public year of one cited source (arXiv v1, blog or model-card release), not an adoption date. The 33 entries split the first 13 of the 14 groups of Table 16.1, without Yang et al. (2024b); vertical position within a group separates sources and encodes no quantity. Row codes refer to the lattice (Tables 14.1 to 14.3). The corpus is not exhaustive, and the readings are interpretations, not evidence for the lattice. In particular, bounded state requires contracting writes ($a < 1$); listing a method does not certify that condition. Exact source dates and their qualifications are kept with the figure's records (Bruce, 2026d).

(Section 0.3.3), the match holds by construction: it shows that the account is consistent, not that it is independently confirmed.

Four controlled deviations tested the blueprint's generative use (Bruce, 2026d), three of them measured on trained models and one, (c), in graph label propagation and synthetic node classification. (a) Propagate-once aggregation, an explicit anchored resolvent solve interleaved with re-seeding blocks, reached parity where filter strength is depth's binding role and lost where layers must restructure the operator, as registered (on routing; it tied on copying); the outcome is directionally consistent but partial, with the vision comparison incomplete and the equivalence between filter strength and depth unassessed. (b) The symmetric-structure encoder is the minimal block above. (c) Conditioning-gated amplification, which sets the filter strength from measured conditioning, was field-tested three times; its dominance condition is sufficient only, and Section 16.1 reports the three tests from their records. (d) Signed, learnable gain shaping helps where task relevance sits at low or negative eigenvalues (Section 2.3; Bruce, 2026i).

The block is the minimal aligned-anchored block, and its ingredients line up with the inventory. Its independent value map is the free content path (Chapter 6). Its $1/L$-scaled residual update keeps the mixing budget bounded, and with it short-path mass (Section 2.6, Chapter 5). Its register tokens are the abstention channel (Chapter 8). Its one symmetric structure map, used where directed routing is not needed, makes every layer reversible. In its decoder form the $\log N$ logit scale is the routing budget (Chapter 9), which in trained language models helped only on interpolated positions (Section 17.4), and separate queries and keys, with the causal mask, supply the arrow (Chapter 10). The full account of the block is left for later work.

At small scale the block exceeded parity. Over the twenty paired runs against parameter-matched ViT-Tiny, single margins ran from 1.37 to 4.34 points, and the 1.8 to 3.9 points quoted above (Bruce, 2026d) are the range of their means over seeds at each dataset and learning rate. The block's deletion record removes one ingredient at a time from a small encoder of the block trained on CIFAR-100, at depths 4, 12 and 24 with five seeds each. Without the identity path the block is at chance from depth 12 on (Section 5.2). Tying the value map to the keys, which also removes 42–49% of the parameters, cost 3.45, 4.62 and 6.16 points of top-1 at the three depths, averaged over seeds. Lowering the value map's rank to $C - 2$, or removing the register, moved the seed mean of top-1 by at most 1.37 points.

16.7 The scaled margin has not yet been replicated

At 22 million parameters, after 90 training epochs, the recorded margin is larger. On Tiny-ImageNet and ImageNet-100, the minimal block's mean top-1 accuracy exceeded ViT-S's by 10.59 and 8.02 points, in 5 of 5 and 5 of 5 paired seeds (exact two-sided sign tests, $p = 0.0625$ and 0.0625, the smallest value that five pairs allow). These are tabulations of recorded results, not a reproduction. The baselines were not tuned, and the recorded DeiT-S differs from ViT-S only by label smoothing. The comparison therefore waits for tuned baselines, and until they return the margin is not evidence of dominance. The record also holds deletions from the block, three seeds each; the one that removes the attention sublayer's identity path is the one Section 5.2 reads.

Pending: Replication against tuned baselines: the minimal block against ViT-S and DeiT-S trained with the DeiT recipe at matched parameters and epochs, with a separate ResNet-50 reference that is untuned and not matched in parameter count.

Pending: Which factor carries the gain: accuracy changes under selected one-factor moves from the block toward a tuned ViT-S and from the tuned ViT-S toward the block, with the anchor varied in strength rather than deleted.

At small scale, resolution transfer is a registered failure: the $\log N$ budget did not carry over to the minimal block. Its factor is one during single-resolution training, so the test measured the evaluation-time rescale (Section 9.6).
Pending: Resolution transfer for the minimal block and the ViT, with and without the $\log N$ budget: top-1 accuracy against evaluation resolution, and at 22 million parameters after the fix.

16.8 What holds in trained networks

Exact under premises checked network by network. Corners under concave prices (Sections 16.1 to 16.4): a theorem about allocations under concave prices; its premise, concave pricing of protections at the level of architectures, is assumed, not measured.

Measured in trained networks. The minimal block's margin at 22 million parameters (Section 16.7): recorded at 22M parameters with 5 paired seeds against untuned baselines; not evidence of dominance until the tuned replication returns.

An interpretation, with nothing to transfer. Adopted changes read as moves in the lattice (Section 16.5): a reading of adopted changes; there is nothing to transfer. The blueprint, the inventory run forward to the smallest block that meets it (Section 16.6): a reading of an assembled constraint set. That reading the past is not evidence (Section 16.5): a statement of scope.

Part V

Where the design can still move

Chapter 17

Abstention and retrieval at long context

Part IV read the standard block as paying for every condition at once. That is a reading under one price list: under it the transformer is the cheapest point, not the only one. This part turns to the room, where a condition is patched rather than met, or paid at a price that another design could avoid (Section 0.3.7). Chapter 17 takes abstention at long context, Chapter 18 bounded memory, and Chapter 19 the machines that pay the same conditions in other currencies. Chapter 20 gives the predictions stated in advance, by which the account can fail.

Chapter 0 states this chapter's two results in brief (Section 0.3.7), with a figure that sets the controlled pilot beside a linear memory and product keys (Figure 0.4): the two-sided law that governs abstention at long context, for a null that absent and present queries share, and the result that decisions on maxima need no length schedule against backgrounds bounded below their threshold. This chapter develops them from Bruce (2026c), Bruce (2026d) and Bruce (2026e). Section 17.1 recalls the two-sided law of Chapter 9. Section 17.2 tests it in trained heads under control and Section 17.3 in a production model, and Section 17.4 in trained vision transformers and language models. Section 17.5 turns from sums to maxima, Section 17.6 gives the rule to use now and the limits of the evidence, and Section 17.7 says what of this holds in trained networks.

17.1 No fixed rule keeps both sides

The first cell of the room is abstention. A softmax row sums to one, so a head cannot abstain independently of its values (Section 8.1), and every sink-logit or headwise-gated head in use, from off-by-one softmax and learned sink logits to value-null registers and headwise output gates, is a gate times an ordinary softmax (Section 8.2); sigmoid weights, which drop the normalization, obey the same two-sided budget (Proposition 9.3). Sections 9.1 and 9.2 proved the two-sided law that governs this cell, and this chapter applies it. With content logits bounded by a fixed τ, no null that matched absent and present queries share, whether fixed, learned or raised by $\log N$, keeps abstention a_* and single-key retrieval w_* beyond $N_* = e^{2\tau}(1-a_*)(1-w_*)/w_*$ background keys (Theorem 9.1). A null that may differ between the two queries buys back at most the abstention penalty, the factor $1-a_*$ in N_*, and never the read's dilution (Proposition 9.2). Multiplying every score, the null included, by $\lambda_N = \log N/\log N_0$ for a head trained at N_0 keys keeps both when the margins are separated (Theorem 9.6), and the margins are real conditions (Theorem 9.8). These theorems bound attention weights under a fixed score law: they do not identify the score law a network learns, and they do not turn an accuracy onset into a weight threshold. The rest of the

chapter asks what trained heads do. That a winning logit needs a margin growing like $\log N$ is known (Nakanishi, 2025; Veličković et al., 2025), and practice already scales logits with length (Chiang & Cholak, 2022; Peng et al., 2024); the two-sided law adds the abstention half, and the tempered rule multiplies the null together with the scores.

17.2 A controlled pilot

The laws of Sections 9.1 and 9.2 describe one head with given scores. A controlled pilot trains such heads and asks whether they still behave as the laws say once their scale, sink and register are learned (Bruce, 2026c). Each head performs lookup-or-abstain. A query is compared with N keys by cosine similarity, with queries and keys isotropic in $\mathbb{R}^{128}$. Half the queries have an exact match, and the head must return its value, one of 16 one-hot classes; the other half have none, and the head must return zero. The output is $o = \sum_j w_j v_j$, trained by squared error, and the logits are $s_j = \tau \cos_j + \beta_0$ with a learned scale $\tau = \tau_{\max}\sigma(\theta)$ bounded by a budget $\tau_{\max} \in \{10, 20\}$, where σ is the logistic function. Seven rules share this geometry:

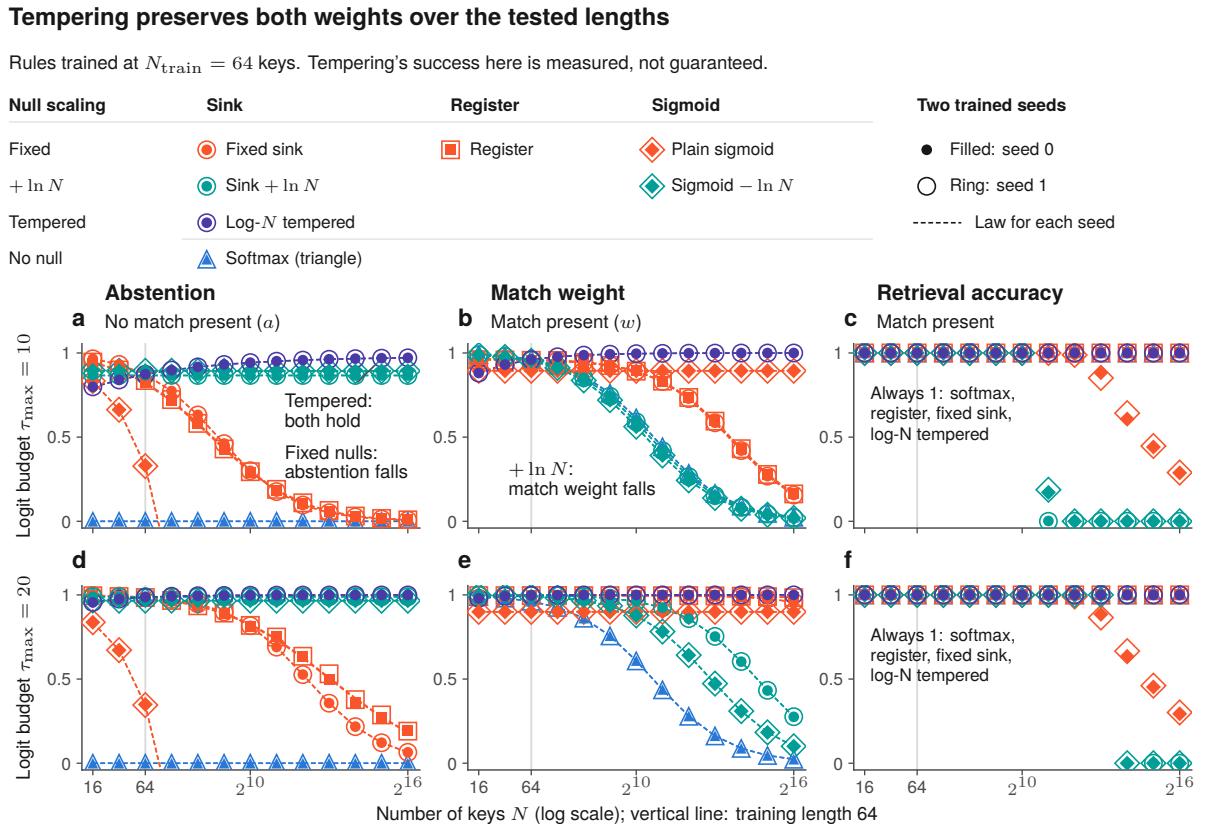
- **softmax**, which cannot abstain;
- a **value-null register**: an extra key with zero value and logit $\tau \cos(q, k_{\text{reg}}) + \beta_0 + b$, whose direction is random;
- a **fixed sink**: a learned null logit b, the form gpt-oss uses (OpenAI, 2025);
- **sink** $+\log N$: the sink logit raised to $b + \log N$, calibrated to the length;
- **sigmoid**: $w_j = \sigma(s_j + b)$, with no normalization;
- **sigmoid** $-\log N$: $w_j = \sigma(s_j + b - \log N)$, the calibrated recipe of Ramapuram et al. (2025);
- $\log N$**-tempered**: the fixed sink with every score, the sink included, multiplied by $\lambda_N = \log N / \log N_{\text{train}}$.

Each rule is trained at $N_{\text{train}} \in \{64, 512\}$ keys under each budget, with two seeds, and evaluated in float64 at every power of two from 16 to 65,536 keys (the protocol is in Bruce, 2026c, Appendix B; all four configurations are in Figure 17.2 and Table 17.4). Since $\lambda_N = 1$ at the training length, the tempered head trains exactly as the fixed sink does and has the same parameters. Tempering acts only at evaluation, so in this pilot it costs no training.

A head abstains on a query when its row mass $\sum_j w_j$ is below $1/2$. It retrieves correctly when the mass is at least $1/2$ and the match's class is the largest entry of o. Two estimands are reported and kept apart. The weights are the abstention $a = 1 - \sum_j w_j$ on absent queries and the match weight w on present ones; a weight's onset is the first length at which its mean falls below $1/2$. The accuracies are the fractions of absent queries that abstain and of present queries that retrieve; an accuracy's onset is the first length at which it falls below 0.9 (Table 17.1).

Figure 17.1 draws both logit budgets at $N_{\text{train}} = 64$, with the law's prediction for each seed's trained head: $\log N$ tempering keeps both weights over the tested lengths, while each other rule loses abstention or the match weight.

Each fixed rule fails at least one side. At $\tau_{\max} = 10$ and $N_{\text{train}} = 64$ (Figure 17.1), the fixed sink keeps retrieval and loses abstention. Its absent-query abstention falls from 0.874 at the training length to 0.097 at 4,096 keys and 0.007 at 65,536 (seed means), and its abstention accuracy fails from 512 keys in both seeds. The value-null register fails the same way, from 128 keys in seed 0 and 256 in seed 1. Both retrieve correctly at every length. At this setting, the calibrated rules do the opposite. Sink $+\log N$ and sigmoid $-\log N$ hold their abstention near its training value (0.866 and 0.893 at 65,536 keys) but lose retrieval from 2,048 keys in both seeds, where the match weight has fallen to 0.42 and 0.39. Softmax never abstains. Plain sigmoid cannot hold its row mass: its absent-query row mass exceeds one from 128 keys in every configuration and seed, and at $\tau_{\max} = 10$,



Plain sigmoid: abstention is negative from $N = 128$, below the plotted range.
Its row mass at 65,536 keys: 689.4/688.8 ($\tau_{\max} = 10$), 671.4/670.7 ($\tau_{\max} = 20$), for seeds 0/1.

Figure 17.1: **Tempering keeps both weights over the tested lengths; every other rule loses one of them.** Seven attention rules, trained at 64 keys, at two logit budgets (rows). Colour says how a rule's null scales with the number of keys N: fixed (orange), raised by $\ln N$ (teal) or tempered together with every score (violet); softmax (blue) has no null. Shape says the mechanism: a sink (circle), a register (square) or a sigmoid (diamond); softmax is a triangle. (a, d) Abstention on queries with no match, $a = 1 - \sum_j w_j$: the fixed nulls lose it as N grows. (b, e) The match weight w on queries with a match: the nulls raised by $\ln N$ lose it. (c, f) Retrieval accuracy, which also requires the row's total weight to be at least one half; softmax, the register, the fixed sink and tempering stay at one throughout. A filled mark and a ring are seeds 0 and 1; dashed lines are the law's prediction for each seed's trained head. Plain sigmoid's abstention turns negative from $N = 128$, below the plotted range. Tempering's success here is measured, not guaranteed: its trained margins over the background do not meet the conditions of Theorem 9.6. The pilot uses isotropic keys in $\mathbb{R}^{128}$, exact matches and 1,024 evaluation queries per length and seed (the protocol is in Bruce, 2026c, Appendix B).

	Softmax	Value-null register	Fixed sink	Sink $+\log N$	Sigmoid	Sigmoid $-\log N$	$\log N$-tempered
$\tau_{\max} = 10$, $N_{\text{train}} = 64$							
learned τ	7.6	9.8	9.8	9.7	6.9	9.8	9.8
abstention fails	$\leq$16	128/256	512	—	64	—	—
retrieval fails	—	—	—	2,048	8,192	2,048	—
$\tau_{\max} = 10$, $N_{\text{train}} = 512$							
learned τ	9.6	9.8	9.8	9.8	1.1	9.9	9.8
abstention fails	$\leq$16	256	1,024	—	64	—	—
retrieval fails	—	—	—	32,768	$\leq$16	16,384	—
$\tau_{\max} = 20$, $N_{\text{train}} = 64$							
learned τ	7.6	16.7/16.6	15.2	15.6	6.9	12.9	15.2
abstention fails	$\leq$16	2,048	8,192	—	64	—	—
retrieval fails	—	—	—	32,768	8,192	16,384	—
$\tau_{\max} = 20$, $N_{\text{train}} = 512$							
learned τ	10.6	19.6	17.8	17.8	2.0	15.0/15.1	17.8
abstention fails	$\leq$16	8,192/16,384	65,536/32,768	—	64	—	—
retrieval fails	—	—	—	—	$\leq$16	—	—

Table 17.1: Pilot accuracy onsets: the first length N at which abstention accuracy (absent queries) or retrieval accuracy (present queries) falls below 0.9, with the learned logit scale τ. Entries show seeds 0 and 1 as "seed 0/seed 1" where they differ; — means no onset up to 65,536 keys, and $\leq$16 means below 0.9 already at the shortest length. Weight onsets, which can differ, are in Table 17.3. The tempered head is the fixed sink's trained head, tempered at evaluation.

$N_{\text{train}} = 64$ it reaches 689 at 65,536 keys, where its retrieval accuracy has fallen to 0.29 (from 0.99 at 4,096). At $N_{\text{train}} = 512$ it does not learn to retrieve at its training length: its learned τ is 1.1 and 2.0 under the two budgets, and its retrieval accuracy there is 0.12 and 0.31. The tempered rule keeps both sides. At $\tau_{\max} = 10$, $N_{\text{train}} = 64$ its abstention does not fall (0.958 at 4,096 keys and 0.971 at 65,536), and none of its weights or accuracies reaches an onset at any length, in any configuration or seed. Table 17.1 gives every rule's accuracy onsets in all four configurations.

A larger budget delays the failures but keeps the trade. Doubling the budget to $\tau_{\max} = 20$ at $N_{\text{train}} = 64$ moves the fixed sink's abstention failure to 8,192 keys, and the calibrated rules' retrieval failures to 16,384 keys (sigmoid $-\log N$) and 32,768 (sink $+\log N$). Training at 512 keys as well moves them further. The calibrated rules then retrieve to 65,536 keys. The fixed sink's abstention accuracy first fails at 65,536 keys in seed 0 and in the seed mean, while seed 1 dips to 0.898 at 32,768 keys; its abstention weight falls below 1/2 at 65,536 keys in both seeds. Only the rules that use the budget move. Softmax and sigmoid learn $\tau \approx 7.6$ and $\tau \approx 6.9$ at $N_{\text{train}} = 64$ under either budget, and their onsets there do not change.

Accuracy and weight are different estimands. The two onsets need not agree, and the pilot shows both directions. Softmax retrieves correctly at every length, although at $\tau_{\max} = 10$ and $N_{\text{train}} = 64$ its match weight falls below 1/2 from 2,048 keys, to 0.024 at 65,536: a correct read needs only that the match's class lead the others, and the background spreads its weight over 16 classes. Conversely, at $\tau_{\max} = 20$ and $N_{\text{train}} = 512$ the register's mean abstention stays above 1/2 to 65,536 keys (0.541 there), while its abstention accuracy fails from 8,192 keys in seed 0 and 16,384 in seed 1. The mean hides a growing share of rows that no longer abstain.

The law tracks the trained heads. The mean-field law of Section 9.2 assumes an independent background, here one of fixed, isotropic keys. With its expectation computed exactly for isotropic

				Supremum margins			Per row			
$(\tau_{\max}, N_{\text{train}})$	Seed	τ	$b-\beta_0$	v_a	v_r	u_r	median v_a	drift	Law error	Failure N
(10, 64)	0	9.78	6.46	1,777	64	0.036	0.73	1.59	0.0018	2.3×10^{12}
	1	9.78	6.46	1,779	64	0.036	0.73	1.59	0.0023	2.3×10^{12}
(10, 512)	0	9.85	7.19	7,314	512	0.070	5.10	1.01	0.0022	8.4×10^{6}
	1	9.85	7.19	7,304	512	0.070	5.10	1.01	0.0022	8.6×10^{6}
(20, 64)	0	15.18	9.32	22,449	64	0.003	0.12	2.47	0.0007	9.4×10^{12}
	1	15.17	9.31	22,439	64	0.003	0.12	2.47	0.0010	9.3×10^{12}
(20, 512)	0	17.80	11.64	241,020	512	0.002	0.48	1.83	0.0012	1.5×10^{14}
	1	17.80	11.65	241,450	512	0.002	0.48	1.83	0.0015	1.5×10^{14}

Table 17.2: The tempered head at its trained parameters. The supremum margins v_a, v_r and u_r of Theorem 9.6 use the background supremum, a key at cosine one; the theorem certifies the head only if all three are at most one. The per-row columns give the median row's v_a at the training length, from its largest background score, and the drift: how far that maximum rises, in logits, between N_{train} and 65,536 keys. The law error is the largest gap between the measured abstention and the stationary power-mean law over the grid. Failure N is where that law puts abstention below 1/2, an extrapolation beyond the grid.

keys and each head's learned parameters, and with nothing fitted, it tracks the trained fixed-sink heads to within 0.002 per seed in abstention (largest error 0.0017) and 0.00022 in match weight, in all four configurations. The six untempered rules have 96 weight sides in all, two for each combination of rule, configuration and seed: abstention and match weight. On the 64 with a measured onset, the measured onset and the law's agree to within one grid step. On the other 32, the law places no onset on the grid either (Figure 9.2, printed with the law in Section 9.2, and Table 17.3).

Tempering works here without meeting its certificate. The tempered rule's trained margins do not satisfy Theorem 9.6. Measured against the background supremum, a key at cosine one, the absent margin is $v_a = 1{,}777$ and 1,779 in the two seeds at $\tau_{\max} = 10$, $N_{\text{train}} = 64$, and between 7,304 and 241,450 in the other configurations; the present margin v_r equals N_{train}, because a background key at the supremum would score as high as the match. Per row the margins look better: at the training length, the median row's largest background score gives $v_a = 0.73$ at $\tau_{\max} = 10$, $N_{\text{train}} = 64$. But the theorem needs the bound fixed before scaling, and the typical row's background maximum rises with N, by 1.6 logits between 64 and 65,536 keys in that configuration, so the per-row reading does not transfer. What describes the tempered heads is the power-mean law of Theorem 9.7, with the background's population power mean in place of its sample one. This stationary law is checked numerically and not proved: evaluated with each head's learned parameters, it reproduces the measured tempered abstention to within 0.0023 over the grid. It predicts that tempered abstention falls below 1/2 near $N = 2.3 \times 10^{12}$ keys at $\tau_{\max} = 10$, $N_{\text{train}} = 64$, near 8.4–8.6×10^{6} at $\tau_{\max} = 10$, $N_{\text{train}} = 512$, near 9.3–9.4×10^{12} at $\tau_{\max} = 20$, $N_{\text{train}} = 64$, and near 1.5×10^{14} at $\tau_{\max} = 20$, $N_{\text{train}} = 512$ (Table 17.2). These lengths extrapolate a law checked only to 65,536 keys; they are not measurements.

The four configurations in full. The two seeds agree closely for most rules, but not for all. Of the 224 printed seed-mean entries, 13 have seeds that differ by more than 0.01, all of them for the register, plain sigmoid or softmax. The largest ranges are 0.151, in the register's learned τ at $\tau_{\max} = 20$, $N_{\text{train}} = 64$; 0.774, in plain sigmoid's abstention at 65,536 keys, a value near -670; and 0.010 for softmax. For the other four rules no entry differs by more than 0.0081. Onsets are

	Softmax	Value-null register	Fixed sink	Sink $+\log N$	Sigmoid	Sigmoid $-\log N$	$\log N$-tempered
$\tau_{\max}=10$, $N_{\text{train}}=64$							
$a<1/2$, measured	≤16	512	512	—	64	—	—
$a<1/2$, law	≤16	360/380	440	—	48	—	2.3×10^{12}
$w<1/2$, measured	2,048	16,384	16,384	2,048	—	2,048	—
$w<1/2$, law	1,600	12,000	12,000	1,500	—	1,300	—
$\tau_{\max}=10$, $N_{\text{train}}=512$							
$a<1/2$, measured	≤16	1,024	1,024	—	64	—	—
$a<1/2$, law	≤16	610	910	—	51	—	$8.4\times10^{6}/8.6\times10^{6}$
$w<1/2$, measured	16,384	16,384	16,384	8,192	≤16	4,096	—
$w<1/2$, law	10,000	12,000	12,000	4,700	≤16	4,000	—
$\tau_{\max}=20$, $N_{\text{train}}=64$							
$a<1/2$, measured	≤16	8,192/16,384	8,192	—	64	—	—
$a<1/2$, law	≤16	8,300/8,600	4,600/4,500	—	49	—	$9.4\times10^{12}/9.3\times10^{12}$
$w<1/2$, measured	2,048	—	—	32,768	—	8,192	—
$w<1/2$, law	1,600	$6.1\times10^{6}/5.3\times10^{6}$	1.6×10^{6}	25,000	—	7,300/7,400	—
$\tau_{\max}=20$, $N_{\text{train}}=512$							
$a<1/2$, measured	≤16	—	65,536	—	64	—	—
$a<1/2$, law	≤16	75,000/78,000	33,000	—	52	—	1.5×10^{14}
$w<1/2$, measured	32,768	—	—	—	≤16	65,536	—
$w<1/2$, law	25,000/26,000	7.1×10^{7}	1.6×10^{7}	2.2×10^{5}	≤16	57,000	—

Table 17.3: Pilot weight onsets: the first length at which the mean abstention a (absent queries) or the mean match weight w (present queries) falls below $1/2$, measured and as the mean-field law predicts from each head's learned parameters. The law's onset is continuous, to two significant figures; the measured one is a point of the grid $16, 32, \ldots, 65{,}536$. Entries show "seed 0/seed 1" where the seeds differ. — means no onset (measured: up to 65,536 keys; law: no root through 10^{30}), and ≤16 means below $1/2$ already at the shortest length. The tempered rule's law is the stationary power-mean law, checked numerically.

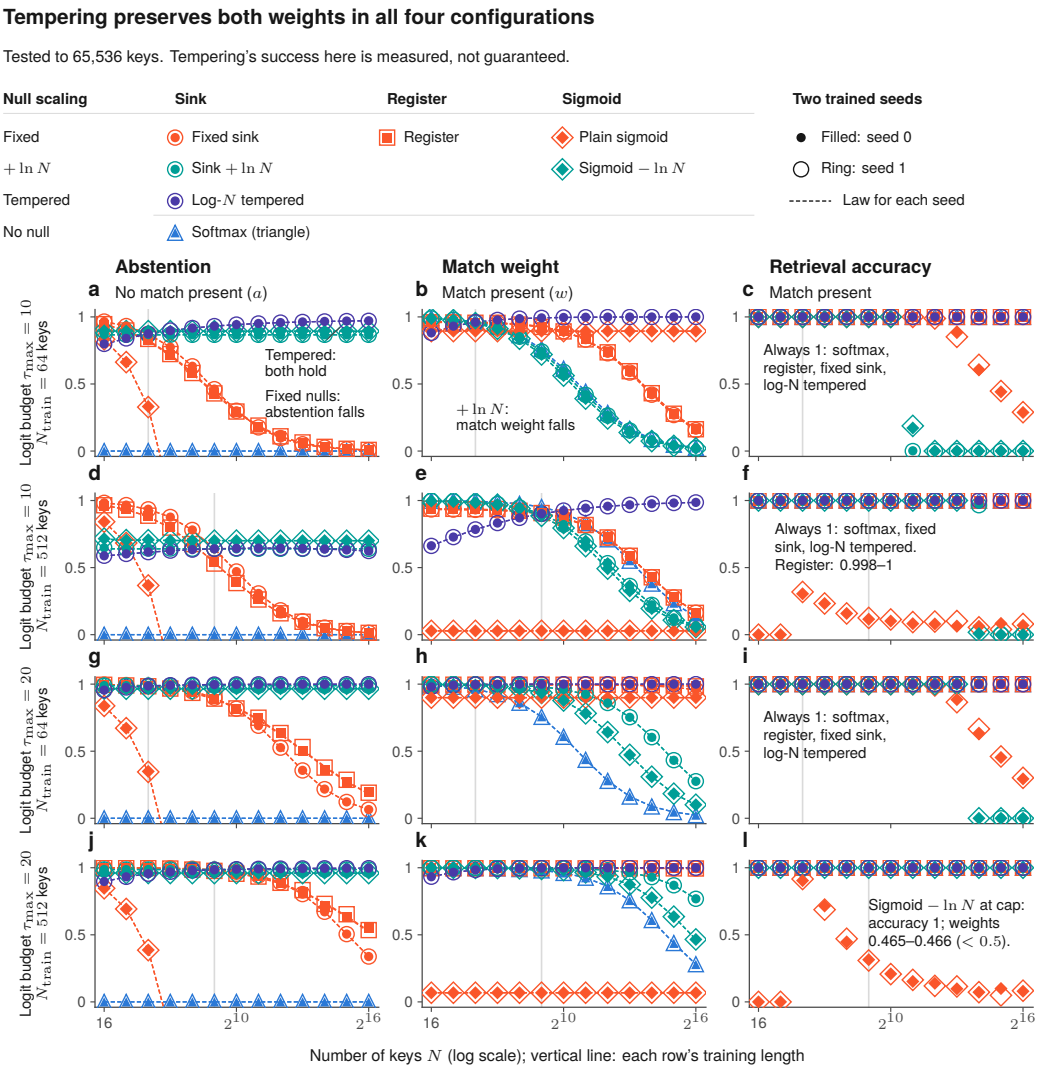
therefore given per seed (Tables 17.1 and 17.3). Table 17.3 gives the weight onsets, measured and as the mean-field law predicts; Table 17.4 gives the seed means of the learned scale, the abstention and the accuracies; and Figure 17.2 draws every rule's weights and accuracy in all four configurations.

Scope. The pilot fixes the feature geometry: keys are isotropic and fixed, matches are exact, and only the scale τ, the offset β_0 and the null logit b are learned. A trained network can also move its background logits with N, which this pilot rules out by construction; gpt-oss's background falls with position (Section 17.3). And since the matches are exact ($c=1$) and the code is noiseless, only one of the two limits of Sections 9.3 and 9.4 binds in the pilot: the limit of the read, not the code limit of the keys.

Other key geometries and training lengths. A registered extension relaxes the geometry, the values and the training length, with the pilot's training and evaluation and five seeds per setting; its design, tables and records are in Bruce (2026c, Section 6 and Appendix B). The same heads are trained at both budgets and $N_{\text{train}}=64$ in the pilot's geometry and seven others: keys drawn around 16 clusters, background keys with a heavy-tailed component along the query, a noisy match at cosine 0.95, 0.80 or 0.60 among four hard distractors, continuous values, and four heads. The rules that sum over the keys keep the pilot's pattern: none without a length input keeps both sides where the law puts a failure inside the grid. The registered outcome is nonetheless partial, because the pattern needs two qualifications, and with near matches at cosine 0.80 and 0.95 at $\tau_{\max}=20$ the register's abstention is neither kept nor lost by the registered thresholds in three cells (Bruce, 2026c). With heavy-tailed keys every rule that sums over the keys loses a side, tempering included.

		Abstention a			Retrieval acc.	Abstention acc.	Seed
Rule	τ	N_{train}	4,096	65,536	4,096 / 65,536	4,096 / 65,536	range
$\tau_{\max} = 10$, $N_{\text{train}} = 64$							
Softmax	7.62	0.000	0.000	0.000	1.00 / 1.00	0.00 / 0.00	0.006
Value-null register	9.84	0.831	0.109	0.009	1.00 / 1.00	0.00 / 0.00	0.011
Fixed sink	9.78	0.874	0.097	0.007	1.00 / 1.00	0.00 / 0.00	0.001
Sink $+ \log N$	9.68	0.868	0.867	0.866	0.00 / 0.00	1.00 / 1.00	0.001
Sigmoid	6.86	0.330	−42	−688	0.99 / 0.29	0.00 / 0.00	0.638
Sigmoid $- \log N$	9.79	0.895	0.893	0.893	0.00 / 0.00	1.00 / 1.00	0.001
$\log N$-tempered	9.78	0.874	0.958	0.971	1.00 / 1.00	1.00 / 1.00	0.001
$\tau_{\max} = 10$, $N_{\text{train}} = 512$							
Softmax	9.58	0.000	0.000	0.000	1.00 / 1.00	0.00 / 0.00	0.000
Value-null register	9.84	0.535	0.161	0.014	1.00 / 1.00	0.02 / 0.00	0.009
Fixed sink	9.85	0.640	0.182	0.014	1.00 / 1.00	0.00 / 0.00	0.000
Sink $+ \log N$	9.85	0.640	0.640	0.640	1.00 / 0.00	1.00 / 1.00	0.000
Sigmoid	1.07	−4.065	−40	−647	0.09 / 0.07	0.00 / 0.00	0.024
Sigmoid $- \log N$	9.87	0.701	0.701	0.701	1.00 / 0.00	1.00 / 1.00	0.001
$\log N$-tempered	9.85	0.640	0.645	0.623	1.00 / 1.00	1.00 / 1.00	0.001
$\tau_{\max} = 20$, $N_{\text{train}} = 64$							
Softmax	7.60	0.000	0.000	0.000	1.00 / 1.00	0.00 / 0.00	0.008
Value-null register	16.63	0.983	0.635	0.194	1.00 / 1.00	0.71 / 0.09	0.151
Fixed sink	15.18	0.986	0.527	0.065	1.00 / 1.00	1.00 / 0.00	0.003
Sink $+ \log N$	15.55	0.989	0.989	0.989	1.00 / 0.00	1.00 / 1.00	0.004
Sigmoid	6.94	0.347	−41	−670	0.99 / 0.30	0.00 / 0.00	0.774
Sigmoid $- \log N$	12.91	0.966	0.965	0.965	1.00 / 0.00	1.00 / 1.00	0.008
$\log N$-tempered	15.18	0.986	0.999	1.000	1.00 / 1.00	1.00 / 1.00	0.003
$\tau_{\max} = 20$, $N_{\text{train}} = 512$							
Softmax	10.58	0.000	0.000	0.000	1.00 / 1.00	0.00 / 0.00	0.010
Value-null register	19.56	0.976	0.887	0.541	1.00 / 1.00	0.95 / 0.56	0.027
Fixed sink	17.80	0.985	0.891	0.338	1.00 / 1.00	1.00 / 0.00	0.002
Sink $+ \log N$	17.81	0.986	0.986	0.986	1.00 / 1.00	1.00 / 1.00	0.000
Sigmoid	2.03	−3.913	−38	−628	0.13 / 0.08	0.00 / 0.00	0.092
Sigmoid $- \log N$	15.05	0.960	0.960	0.960	1.00 / 1.00	1.00 / 1.00	0.004
$\log N$-tempered	17.80	0.985	0.994	0.997	1.00 / 1.00	1.00 / 1.00	0.002

Table 17.4: The four pilot configurations in full: seed means of the learned scale, of the abstention a at the training length, 4,096 and 65,536 keys, and of retrieval and abstention accuracy at 4,096 and 65,536 keys, with each rule's largest range between the two seeds over these entries. Plain sigmoid's abstention is negative because its row mass exceeds one.



Plain sigmoid: abstention is negative from $N = 128$. Row mass at 65,536 keys (seeds 0/1):
(10, 64): 689.4/688.8; (10, 512): 648.3/648.3; (20, 64): 671.4/670.7; (20, 512): 628.8/628.9.

Figure 17.2: **Tempering keeps both weights across all four configurations; other rules need not fail within the tested lengths.** Rows vary logit budget $\tau_{\max}$ and training length N_{train}; columns show absent-query abstention $a = 1 - \sum_j w_j$, present-query match weight w and retrieval accuracy. Colour says how a rule's null scales with the number of keys N: fixed (orange), raised by $\ln N$ (teal) or tempered together with every score (violet); softmax (blue) has no null. Shape says the mechanism: a sink (circle), a register (square) or a sigmoid (diamond); softmax is a triangle. Seven rules, two seeds, 13 lengths (16–65,536 keys), 1,024 queries per length and seed. A filled mark and a ring are seeds 0 and 1; dashed lines are the law's prediction for each seed's trained head. Plain sigmoid's 80 negative abstention values are off scale. At $(20, 512)$, both calibrated rules retain accuracy one, but sigmoid $-\ln N$ has match weight below one half. Tempering is empirical, not certified; $\lambda = \ln N / \ln N_{\text{train}}$ applies below training too. Absent queries have N background keys, present queries $N - 1$; Theorem 9.1 uses N in both.

The calibrated rules lose retrieval, as predicted, and keep their abstention accuracy, but their mean abstention also falls with N, by 0.034 to 0.058, beyond the registered tolerance of 0.02. Tempering loses abstention accuracy from 256 to 1,024 keys. Its own law, given the heavy-tailed background, also puts the mean abstention below 1/2 inside the grid, at 370 to 600 keys, though earlier than the measured mean falls below it and outside the regime in which the law is quantitative. With clustered keys at $\tau_{\max} = 10$, the tempered heads' mean abstention stays above 0.64 to 65,536 keys, and the law puts its fall below 1/2 beyond the grid; yet their abstention accuracy falls below 0.9 inside the grid, to between 0.71 and 0.77 at 65,536 keys, in all five seeds. With isotropic keys, training on lengths drawn from 64 to 1,024 delays the fixed sink's failure without removing it: its abstention accuracy fails from 1,024 keys instead of 512 at $\tau_{\max} = 10$, and from 32,768 instead of 8,192 at $\tau_{\max} = 20$; with clustered or heavy-tailed keys at $\tau_{\max} = 10$ it does not delay the failure. A sink whose null logit adds $c \log(N/64)$, with c learned, keeps both sides only at the larger budget: in all five seeds with isotropic keys, where it learns $\hat{c} = 0.95$ to 1.06, a $\log N$ schedule, and in two with clustered keys ($\hat{c} = 0.63$ to 0.66). At the smaller budget it learns $\hat{c} \leq 0.31$ and loses abstention, and with heavy-tailed keys it loses both sides. Trained on the same lengths with heavy-tailed keys, the calibrated sink's mean abstention falls by 0.09 to 0.22 over the grid, and at the smaller budget it loses abstention in three of five seeds. Tempering applied in training loses abstention with heavy-tailed keys, from 512 to 2,048 keys, and with clustered keys at the smaller budget in four of five seeds. The mean-field law, evaluated with each geometry's own background moment, agrees with the measured weight onset in 1,113 of the 1,129 scored comparisons. For onsets inside the grid, agreement allows one grid step in N from the first grid point at or above the predicted onset; censored onsets have separate comparison rules. All 16 misses are the register with near matches; another 1,151 comparisons are not scored, because the law is outside its concentration regime there or neither onset falls on the grid. Every geometry passes at least 91% of its comparisons over both budgets together, but at $\tau_{\max} = 20$ near matches at cosine 0.80 and 0.95 pass only 20 of 25 and 25 of 29, below the registered 90%, so there the mean-field law is not quantitative. Tempering is certified in no geometry. Of the 120 labelled cases, 114 have their abstention tracked by the stationary law of a concentrated background to within 0.05 at every length. The remaining six are empirical successes: they keep both sides without that explanation. With heavy-tailed keys tempering has no label, because it fails. The law tracks failures as well as successes: at the smaller budget 29 of those 114 cases lose abstention, 20 of them with four heads and the rest with clustered keys, and in 15 more, with near matches, the head did not abstain even at the training length. Two heads that decide by reading the largest score are scored separately. The key check, which reads it for both decisions, keeps both sides in every geometry except heavy-tailed keys, where it loses both, its abstention failing at the wall its extreme-value law predicts; the max gate, which reads it to abstain but routes over every key, also loses both with heavy-tailed keys and keeps both in the other geometries, except that it loses retrieval with the weakest near match at the smaller budget. The need for a budget that grows like $\log N$ is therefore a property of rules that sum over the keys, against backgrounds bounded below the threshold; when the background's maximum grows with N, as with heavy-tailed keys, the heads that read it fail too (Sections 9.4 and 17.5).

17.3 A production model

gpt-oss-20b (OpenAI, 2025) learns one sink logit b_h per head, so it carries the fixed-sink half of the law in a production model. At query position i a head's sink log-odds are exactly

$$y_h(i) \;=\; b_h - \mathrm{LSE}_{j \in K(i)}\, s_{ij} \;=\; b_h - \log n(i) - m_h(i),$$

where $K(i)$ is the set of $n(i)$ visible keys and $m_h(i)$ is their mean background level, the logarithm of the mean of $e^{s_{ij}}$. If the background is stationary, y_h falls against $\log(i+1)$ with slope -1; a network that lowers its background logits like $\log n$ holds it flat (Section 9.2). The model alternates 12 dense layers, where $n(i) = i + 1$, with 12 banded layers whose window holds 128 keys, where $n(i) = \min(i + 1, 128)$. The mechanism is the same and only the key count differs, so the banded layers are a built-in control: beyond the window they should stay flat, and while it fills they should fall like the dense layers. Wang (2026) reports the model's sink mass averaged over positions, 0.48 in its banded layers and 0.34 in its dense ones (their Table 14); the test here asks how the sink's log-odds change with position.

The test was prespecified, not registered; its documents, the freeze of its analysis and its selection of the sink-using heads are in Bruce (2026c, Section 9 and Appendix C). For each sink-using head, we fit the document-mean log-odds on $\log(i+1)$. The three legs apply the following prespecified criteria to the median slopes:

- dense layers at positions $[256, 4096)$, with a median slope of at most -0.5;
- banded layers at positions $[256, 4096)$, with a median slope strictly within ± 0.2;
- banded layers while the window fills, at positions $[8, 128)$, with a median slope of at most -0.5.

All three holding would confirm the law. The dense leg failing while the two banded legs hold would mean that gpt-oss compensates. Anything else is inconclusive. Before any leg is read, a control requires the natural documents' mean next-token loss to beat that of the same documents with their tokens shuffled by more than 1 nat in each of the two corpora, books and code, so that a broken load cannot pass.

The test is undecided by its prespecified rule; descriptively it leans toward compensation. Every implementation and fidelity check passes in every cell: the forward pass is bit-identical to the model's own, and the sink probabilities from the float32 log-odds agree with the bfloat16 path's to median differences of at most 2.6×10^{-4}, against a tolerance of 10^{-3} fixed in advance. The code corpus passes the control, with a gap of 3.87 nats, but the books corpus does not. Its gap is 0.672 nats: 5 of the 32 books have a negative gap (their loss is lower shuffled than in natural order), and 11 exceed 1 nat. The natural books text itself scored 7.06 nats per token, against 3.39 on code, far above what a model of this size reaches on ordinary prose. That leaves the prespecified outcome undecided. An audit of the text and tokens of all 64 documents found no loading defect, and it does not explain why the books control fails. The slopes would not have decided it either (Table 17.5). The dense leg fails, so the confirmation branch is not reached. The banded leg beyond the window holds, but the banded leg while the window fills fails, so the compensation branch is not reached either. Books and code agree (-0.20 and -0.19 on the dense leg). Selecting heads on half the documents and fitting on the other half moves no median by more than 0.012. Shuffling the tokens leaves the slopes about as flat (-0.15, 0.02 and -0.15).

Table 17.5 gives the three legs as measured, and the failed books control makes all three of its rows descriptive.

What the measurement shows. Descriptively, the sink log-odds fall far more slowly than a stationary background allows, in the dense layers and in the banded layers while the band fills. In dense layers the mean background level $m(i)$ drops by about 3.4 nats between positions $[1, 10)$ and $[1000, 4096)$ of the books (3.6 in code), and the implied background mass e^{m-b}, measured per key relative to the sink, falls 30- to 35-fold. The first token takes only 3–4% of the content attention at those positions, so the flatter slope comes with a falling background, not with the first token absorbing the row. The prespecified design did not decide which mechanism lowers the background. The captured rows split the growth of the background log-sum exactly, head by head. Between 256 and 4,096 keys, both contributions are measured per e-fold of the key count. Adding the far keys to

Leg	Positions	Prespecified	Median slope	95% interval	Heads
Dense	$[256, 4096)$	≤ -0.5	-0.202	$[-0.220, -0.173]$	700
Banded	$[256, 4096)$	$\lvert\cdot\rvert < 0.2$	-0.019	$[-0.030, -0.008]$	759
Banded, filling	$[8, 128)$	≤ -0.5	-0.222	$[-0.243, -0.202]$	759

Table 17.5: gpt-oss-20b: slopes of the document-mean sink log-odds on $\log(i+1)$, pooled over the 64 natural documents. Medians over sink-using heads, with 95% intervals from 2,000 bootstrap resamples of documents, stratified by corpus. The failed books control makes all three rows descriptive.

the row raises its log-sum by only 0.35 to 0.56, against 1 for exchangeable keys. The log-sum of a fixed-size window of the most recent keys drifts by -0.17 to -0.31 as the query moves on. These are medians over the 700 sink-using dense heads of each corpus and condition. The captured log-sums reproduce the measured log-odds to within 6×10^{-7}. At the start of that range the window is the whole row, first token included; at its end the first token lies outside it. A distance kernel and a drift of the scores with position both hold the background down, the form that compensation through position takes (Section 9.2). The reading is descriptive: the prespecified test did not reach its compensated branch, and nothing here says whether the pattern would survive a passing control.

Figure 17.3 draws the curves behind these slopes, with the stationary-background reference and the control's loss gaps.

What follows. The production-model test neither confirms the fixed-sink law in gpt-oss nor, by its prespecified branches, shows that gpt-oss compensates. It is also no evidence against the identity above, whose slope of -1 requires a stationary background; gpt-oss's background is not stationary in position. Claims about the fixed-sink law in trained models therefore rest on the controlled pilot (Section 17.2) and the trained models of Section 17.4.

The sink length law is also one of the predictions stated in advance. Section 20.1 enters it with the others, with this outcome and the form measured in trained 125M models, and Section 20.2 says what each outcome of this test would have changed.

17.4 Trained vision transformers and 125M language models

Trained vision transformers. The first trained result is a subset in vision: ViT-S/16 trained on ImageNet-100 at 224 px with five attention rules (softmax, value-null registers, a learned sink, an output gate, and the gate with a learned sink), three seeds each, every model evaluated at 224, 320, 448 and 640 px, which multiplies the number of patch tokens by 1, 2.0, 4.0 and 8.2. Every resolution above 224 px here is the 224 px crop upsampled bicubically, with the position embeddings interpolated bicubically, so it adds tokens but no image content. Its one registered prediction failed: it held that the output gate, alone or with a tempered sink, lowers both the content-sink mass (the attention absorbed by patch tokens that act as sinks) and the ratio of maximum to median absolute activation below softmax's, in every seed at 224 px, and while content-sink mass fell in all three seeds, the activation ratio rose in one seed for both variants; that ratio ranges from 25 to 31 across softmax's own seeds, the outliers it measures are mild in these models, and W8A8 quantization moves any variant's seed-mean accuracy by at most 0.2 points. The other resolutions are reported without a gate, so what follows is descriptive. Tempering the trained sink at evaluation, multiplying every score, the sink's included, by $\log N / \log N_{\text{train}}$ (the schedule of Theorem 9.6), leaves accuracy at 224 px unchanged by construction and raises it by 0.9, 2.0 and 3.5 points at 2.0, 4.0 and 8.2 times the training token count, and by at least 3.1 points in every seed at 8.2 times; the untempered

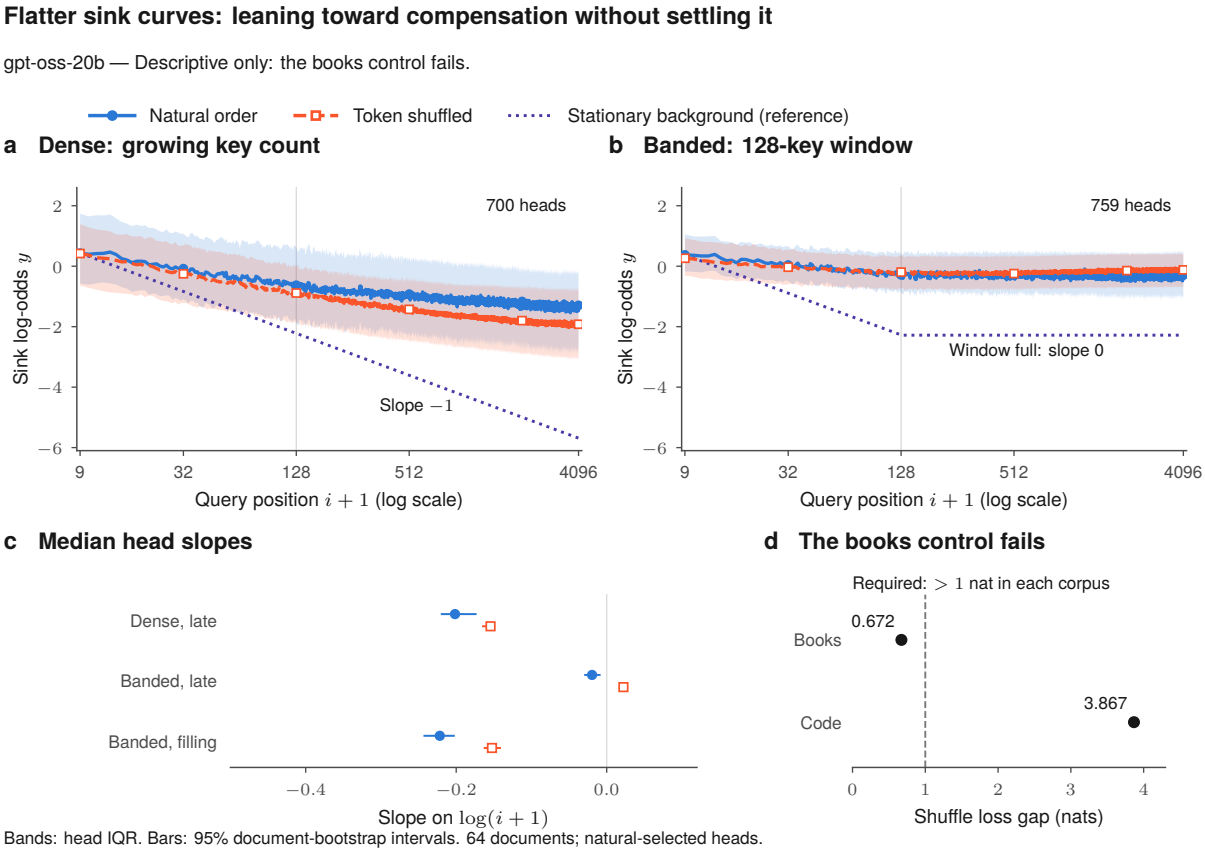


Figure 17.3: **gpt-oss curves depart from the stationary-background reference; the pre-specified test is undecided, though descriptively the slopes lean toward compensation.** (a,b) Median and interquartile range of document-mean sink log-odds across sink-using heads (counts shown), selected on 64 natural documents, 32 per corpus, and reused for the shuffled condition; positions $i \in [8, 4096)$. Dotted references are $y(8) - \log[n(i)/n(8)]$, anchored to each natural median at $i = 8$, with $n(i) = i + 1$ (dense) or $\min(i + 1, 128)$ (banded). (c) Median per-head slopes, not slopes of the median curves: late positions $[256, 4096)$ and filling positions $[8, 128)$. Bars are saved 95% intervals from 2,000 corpus-stratified document-bootstrap resamples with head reselection. (d) Shuffled-minus-natural next-token loss gaps; the books gap falls below the required 1 nat. All measurements are descriptive: neither confirmation nor demonstrated compensation.

sink, the output gate and the value-null registers each stay within 0.6 points of softmax at 640 px. A second evaluation of the same fifteen checkpoints, specified and frozen before any of its numbers existed, added accuracy at 896 px, 16 times the training token count, and softmax tempered the same way with no sink. Both predictions about the extrapolation held: at 896 px tempering raised accuracy by at least 3.7 points in every seed of the learned sink and the gated sink, against a prespecified threshold of 2.0 points, and the attribution test, fixed in advance at 640 px, came out in its branch for log-N scaling alone: tempering the sink gained 3.48 points and tempering plain softmax 4.16, a difference of -0.68 points (95% interval $[-1.30, -0.04]$). The accuracy gain therefore comes from scaling the scores with $\log N$, which the sink's tempering includes, and the sink is not needed for it; this is not a claim that sinks have no other use. These are three seeds of one model size on one dataset, and the confirmation reuses the same checkpoints and validation

images, so it adds measurements, not seeds or images; it is prespecified, not publicly registered; so is every other registration in this section, each frozen before its run. A registered re-evaluation of nine of these checkpoints (softmax, the learned sink and the gate with a learned sink, three seeds each) summarized the CLS rows' attention at 224 to 896 px, untempered and tempered, on two input routes: the upsampled crops above, and the original images resized to each resolution (Figure 17.4). On the upsampled crops the background grows as dilution predicts: the log-sum of the CLS rows' content logits rises with slope 1.01 to 1.04 in $\ln N$, against 1 for the stationary law, while the top logit rises far more slowly (slope 0.25 to 0.30) and the rows' entropy rises at the dilution's rate (slope 0.96 to 0.98), in all 9 checkpoints. Tempering does not restore the trained attention: in the last four layers, where regions, entropy and persistence are compared, it moves the rows' attention further from its 224 px form, region by region (total variation up by 0.027 to 0.045 in 9 of 9 checkpoints), and undoes only 17% to 26% of the entropy's rise. The per-image gain does not track whether the region a row preferred at 224 px, or one of its eight neighbours, keeps its top key, so why tempering helps these models is open. The gain itself holds on resized originals: 4.87, 4.65 and 4.54 points at 896 px, against 4.99, 4.87 and 4.99 on the upsampled crops, although most originals are themselves upscaled at that size. Per head, the learned sink's log-odds fall at -0.96 per unit of $\ln N$, near the stationary law's -1, while with the output gate added to the sink abstention holds level: a null in the scores competes with a background that grows like N, and a gate does not. The corresponding prediction for trained language models, that the tempered rules degrade least with length, failed.

Trained language models. The harness's language-model leg trained a 125M-parameter decoder with each of 11 attention rules, 3 seeds each, at 1,024 tokens, and evaluated it to 8,192 tokens, with YaRN (Peng et al., 2024) as the registered primary positional treatment. Its lead prediction failed (Figure 17.5a–c). On the same checkpoints, tempering on top of YaRN, which already carries a log-length temperature, raised long-context degradation against softmax. The perplexity ratio D divides a checkpoint's long-context perplexity by its perplexity on the same books' training-length prefixes. At 8 times the training length, this ratio exceeded softmax's by $+0.27$ for the tempered learned sink (95% interval $[+0.21, +0.34]$) and by $+0.18$ for the tempered gated sink ($[+0.14, +0.22]$). Of that excess, tempering itself, against the same checkpoint untempered, added 0.10 to 0.17 in each seed. Tempering also did not hold abstention within the registered bound. None of the probe's retrieval heads met the separated margins that Theorem 9.6 needs, so its guarantee was not exercised. Fixed sinks behaved differently on text and on the abstention probe. On text, the sink log-odds of off-by-one softmax and of a learned sink fell with slopes -0.14 and -0.12 in $\log(i+1)$ inside the training length (-0.07 for the sink plus $\log n$), and -0.29 and -0.28 beyond it with YaRN, against -1 for a stationary background, and as shallow as gpt-oss's dense layers (-0.20 and -0.19, Section 17.3); natively, beyond the training length, they fell at about -6.2. On the probe, with YaRN, they kept about a tenth of their null mass at 8 times the training length, from 0.028 to 0.003 and from 0.059 to 0.007. The length-calibrated sink did not follow its pilot pattern: with YaRN it lost abstention (a change of -5.2 in log-odds from the training length to eight times it) and kept its match weight (-0.13), while with native positions it kept abstention and lost its match weight, as in the pilot. Every rule that gives attention a null removed the first-token sink (0.039 of the attention under softmax, about 0.003 or less with a null) and cut massive activations by 67% to 82%; a signed gate without a null did not (its 7% cut has an interval that spans zero). Sigmoid aside, every rule with a null also lowered perplexity at the training length, by 1.2 to 2.5 percent; sigmoid attention raised it by 2.8 percent. With YaRN, softmax and the output gate degraded least with length (D of 1.22 and 1.23 at 8 times). At this scale no retrieval head reached the registered identification threshold (largest weight 0.247 against 0.3), and no rule abstained behaviourally on the probe, so these abstention results describe attention weights, not behaviour. Trained models at 350M parameters are not yet measured.

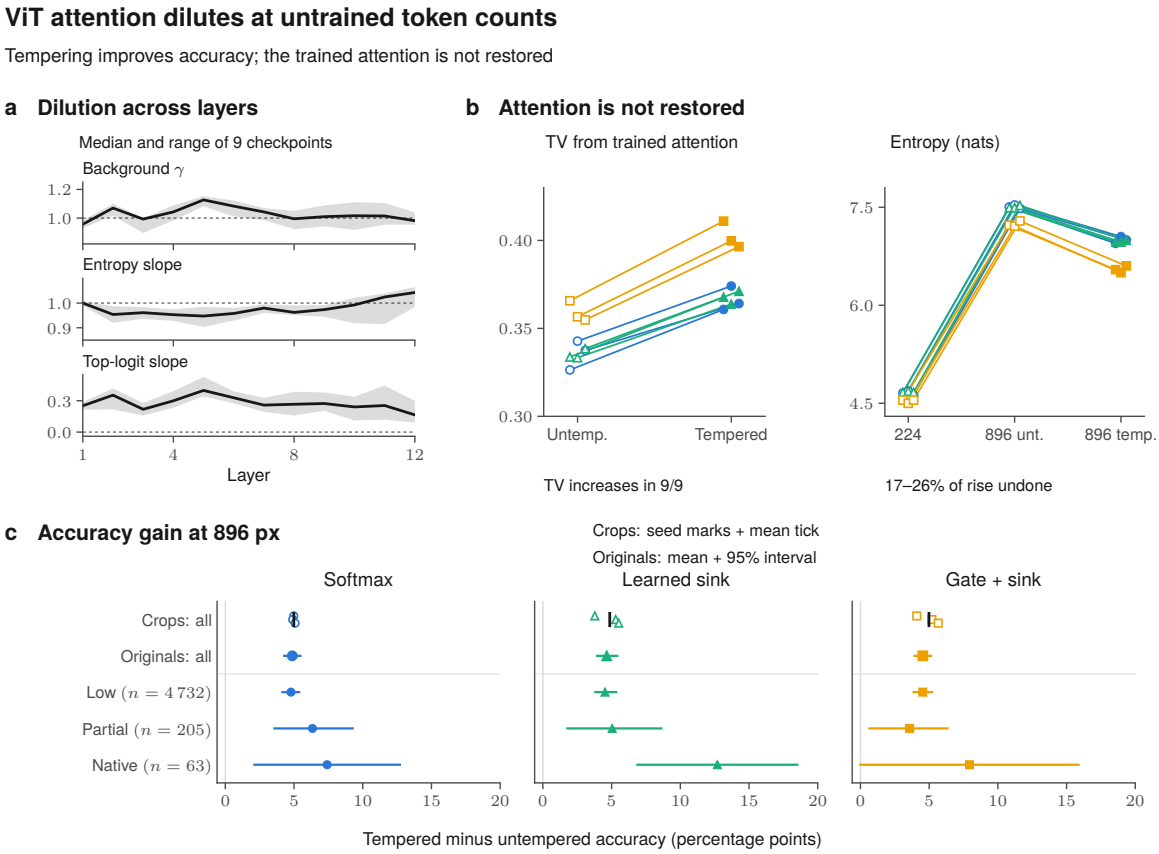


Figure 17.4: **At untrained token counts, ViT attention dilutes, and log-N tempering raises accuracy without restoring the trained attention.** A registered re-evaluation uses 9 checkpoints (3 rules × 3 seeds), one ViT-S/16 size on ImageNet-100. (a) Layer profiles of the registered untempered CLS slopes against $\ln N$ (patch count) over 224–896 px: median across checkpoints and full range; the background log-sum and entropy have stationary-law reference 1, the top logit reference 0. (b) Paired checkpoint summaries, CLS layers 9–12 on upsampled crops: the registered total-variation (TV) comparison with the 224 px regional attention, and descriptive entropy in nats. Open marks are untempered; filled marks are tempered. TV rises in 9 of 9; tempering undoes only 17–26% of the entropy rise (descriptive). (c) Tempered minus untempered top-1 accuracy at 896 px: the registered overall comparison on resized originals and descriptive crop/detail comparisons. For resized originals (O), bars are the recorded 95% paired hierarchical bootstrap intervals, resampling seeds then matched images. For upsampled crops (U), the three small marks are seed gains and the vertical tick their mean; no image-bootstrap interval is available in these summaries. The lower three rows give original-detail strata, with images per seed printed at left: low is below half, partial is at least half but below full, and native is full linear detail. These strata are descriptive. Accuracy uses 5,000 validation images; row summaries use the CLS rows of the first 1,000. The registered dilution and bounded-top/entropy tests held, restoration and gain–persistence tests were falsified, and leader persistence was inconclusive, meeting neither its held nor falsified threshold. Why tempering helps remains open. Most resized originals are also upscaled at 896 px: the median shorter side is 375 px, and 95% carry less than half the native detail.

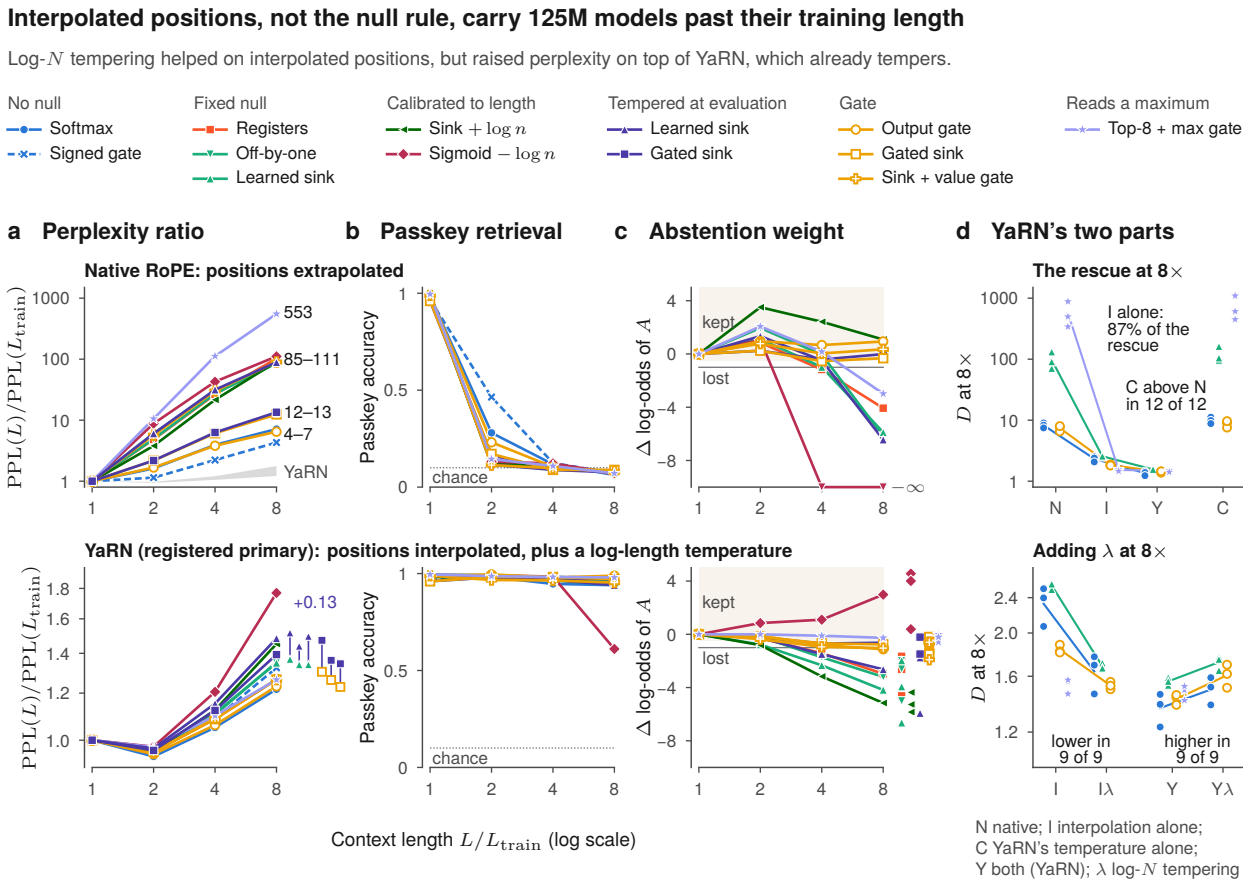


Figure 17.5: **Interpolated positions, not the null rule, carry 125M language models past their training length, and tempering every logit by $\log N/\log N_{\text{train}}$ on top of YaRN, which already carries a log-length temperature, did not keep abstention and raised long-context perplexity.** Lines join the means of 3 seeds per rule. (a) The perplexity ratio D on 95 PG-19 books; the grey band (top) spans every rule with YaRN. Tempering the same checkpoints raised D at 8× from 1.35 to 1.48 (learned sink) and from 1.27 to 1.39 (gated sink), by 0.10 to 0.17 in every seed (bars right of 8×). (b) Passkey accuracy; chance is 0.1. (c) The change from 1× in the log-odds of A, the attention mass a rule withholds from the context on the probe's absent queries (its null's mass, sigmoid's row deficit, or the share a gate closes); lines average each seed's change, as registered. Shading and line mark the registered bounds for keeping it (at least −0.5) and for losing it (at most −1), read under YaRN; marks right of 8× are seeds. At 125M no retrieval head reached the registered identification threshold (largest weight 0.247 against 0.3), so these curves are attention weights, not behaviour. Softmax and the signed gate have no null and are omitted. (d) A prespecified re-evaluation tests four rules' checkpoints at 8× on the first 24 of those books (marks are seeds). Interpolation alone (I) gives 87% of YaRN's rescue in $\log D$ (median; 70% to 99% by rule); the schedule of Theorem 9.6 (λ) lowers D on interpolated positions in 9 of 9 checkpoints and raises it on top of YaRN in 9 of 9; YaRN's temperature alone (C) raises it above native in 12 of 12. Models at 350M are not yet measured.

Learned nulls across training lengths. In the tested byte-level language models with no positional encoding and a learned null, training does not hold abstention fixed as the context grows. The test trained 25 models of 594,320 parameters for 1,500 steps at 5 lengths from 64 to 1,024 tokens, 5 seeds each. The learned null, relative to the mean content logit, does not rise with the training length (slope −0.14 per e-fold, 95% interval [−0.18, −0.10]), and the log-odds of abstaining fall at −0.93 per e-fold of the training length ([−0.98, −0.88]), close to the fixed-null law's −1. Within each model, at fixed weights, they fall at −1.04 to −1.01 per e-fold of the context in the models trained at 128 tokens or more (−1.31 at 64 tokens), as medians over the heads of the five seeds trained at each length. The measured within-model slopes are close to the stationary law's prediction in Section 9.2, except in models trained at the shortest length. These are small models, and the result concerns abstention, not retrieval.

Positions and nulls in small byte models. The same byte-level model, trained at 256 bytes with three positional schemes (none, rotary, rotary on a quarter of each head) and four null rules (none, a learned null, a learned null plus $\ln n$, an output gate), 5 seeds each, separates the positional channel from the null. Without positional encoding the background grows like n under every rule (γ 1.08 to 1.13 inside the training length), and a fixed null's log-odds fall by the stationary law (slope −1.08, 95% interval [−1.09, −1.08]). Adding $\ln(n/N_{\text{train}})$ to the null at evaluation then holds its odds level beyond the training length (slope −0.02, against −1.02 without it) and moves the loss by less than 10^{-4} bits per byte. With rotary positions the background grows like $n^{0.12}$ to $n^{0.19}$ with no null, an output gate or a fixed null alike, so a fixed null's slow decay is the kernel's default. Only a null that itself adds $\ln n$ raises the background growth, to $n^{0.97}$ (difference 0.78, [0.68, 0.89]), and those heads learn less local attention: the median head's top key is among the last 8 in 46% of rows, against 82% with a fixed null. Beyond the training length, without positional interpolation, the heads that held content furthest below a fixed null raise their content logits most (Spearman 0.87, [0.78, 0.92]), and a fixed null's length penalty exceeds plain softmax's by 0.12 and the output gate's by 0.11 bits per byte (5 and 4 of 5 seeds). Whether rotating only a quarter of each head reduces this length penalty is undetermined: it did so in 2 of 5 seeds. These are small models on one corpus, at one training length.

Separating interpolation from temperature. YaRN combines position interpolation (Chen et al., 2023) with a log-length temperature. A prespecified re-evaluation of the 125M checkpoints (softmax, the learned sink, the output gate and the head that reads maxima; 12 checkpoints) applied YaRN's two parts one at a time (Figure 17.5d). Interpolating the rotary frequencies alone carries 87% of YaRN's reduction in the log perplexity ratio at 8 times the training length (median; 70% for softmax, 99% for the head that reads maxima). YaRN's temperature adds to it beyond the training length in 12 of 12 checkpoints, and alone it produces the in-range shift in the sink's log-odds (regressed on the native log-odds below the training length, the shift has slope 0.39 to 0.40, against −0.03 for interpolation). In these models sharpening helped only on interpolated positions. The schedule of Theorem 9.6 on interpolated positions lowers the loss beyond the training length in 9 of 9 checkpoints, by 0.20 to 0.50 nats per token, and raises softmax's passkey accuracy at 8 times from 0.65 to 0.83. Stacked on YaRN, which already tempers, it raises the loss in 9 of 9 (by 0.08 to 0.18 nats per token). YaRN's temperature alone, without interpolation, worsens long-context perplexity in 12 of 12 checkpoints. The harness's tempered arms were evaluated under YaRN, whose temperature also multiplied their sink logit, so every score carried one multiplier, as Theorem 9.6 assumes; their failure was over-sharpening. On these checkpoints YaRN's own temperature beats the schedule of Theorem 9.6 beyond the training length in 9 of 9. The passkey gain, the count for the temperature alone and this last comparison were not among the re-evaluation's prespecified predictions and are descriptive.

Pending: What dropping sum-to-one costs at depth: loss-spike incidence and the per-block Jacobian

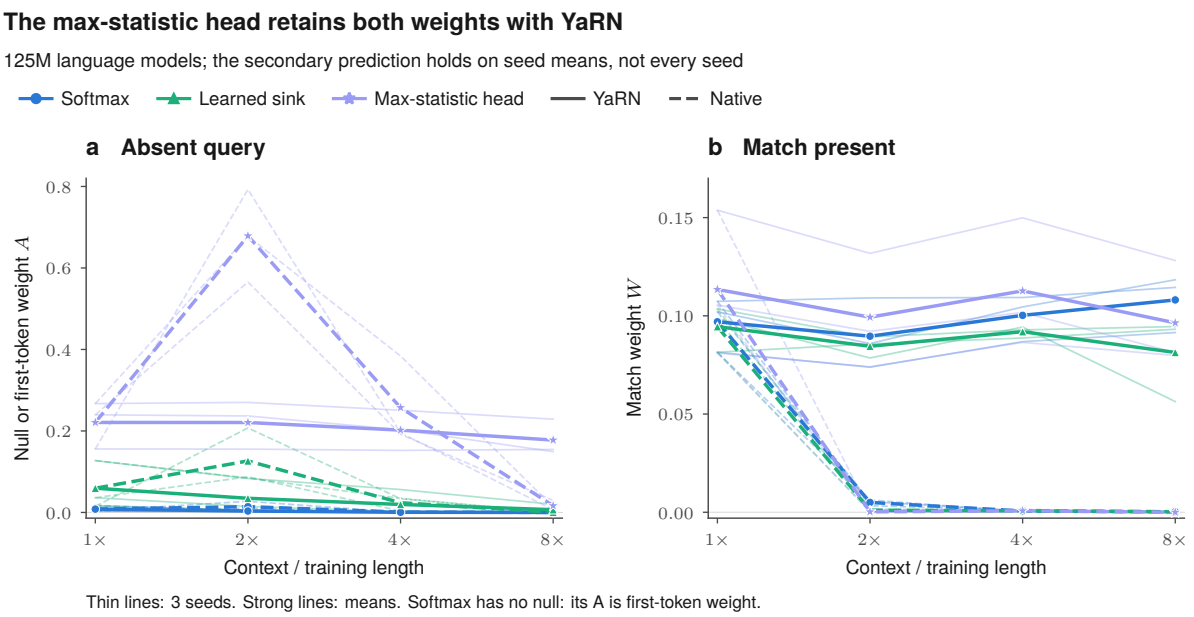


Figure 17.6: **The max-statistic head retains both probe weights with YaRN, while native extrapolation fails.** (a) Absent-query weight A: null mass for the learned sink and max-statistic head, first-token attention for softmax, which has no null. (b) Present-query match weight W. Solid lines use YaRN; dashed lines use native positions. Strong lines are seed means; thin lines show all 3 seeds, without a confidence interval. The models have 125M parameters and trained at 1,024 tokens. Each point averages 64 probes over 3 fallback heads; no head reached the retrieval-identification threshold, so these are attention weights, not behavioural retrieval. The registered secondary prediction holds on mean per-seed log-odds changes from training length to 8×: $\Delta \operatorname{logit} A = -0.27$ and $\Delta \operatorname{logit} W = -0.17$, both above -0.5, but it holds jointly in only 2 of 3 seeds. The max head's rescue is due to positional interpolation (Bruce, 2026c); it does not establish a guarantee beyond the tested lengths.

norm against depth, at 12, 24 and 48 layers, for softmax, off-by-one, output-gated and sigmoid attention, with and without a hybrid norm.

A head that reads maxima. The harness also trained the head of Section 17.5, which routes to its top few keys and abstains when its largest score falls below a learned threshold. With YaRN, it was the only rule whose null and match weights both changed by less than half a nat of log-odds out to eight times the training length, so the registered secondary prediction held on the seed means, though not in every seed (Figure 17.6); five rules, softmax among them, kept the match weight alone better, but each lost more than half a nat on its null or first-token weight. Without YaRN it failed as the matching key fell far down the ranking and the background maximum grew with length, the evidence wall the theory names (Section 9.3).

Pending: Where trained machines sit relative to their walls, the evidence and gain walls of Section 9.3 among them, beyond the first readout on small trained models (Section 19.4).

17.5 Maxima instead of sums

The two-sided law is a law of sums: its null competes with a sum over the N keys, at a fixed logit scale, and matched absent and present queries share it. A rule whose gate and routing read maxima rather than sums, such as a head that routes to its top few keys and abstains when its largest score falls below a learned threshold, is predicted to need no length schedule at a fixed logit scale, against backgrounds whose maximum stays below its threshold at every length. The derivation and the constructive tests belong to Bruce (2026e), and the prediction was fixed in the trained harness's protocol before the run; its result is in Section 17.4. In a trained model that condition belongs to the positions, not to the rule.

Chapters 8 and 9 print the derivation and the constructive tests. A test that reads the largest score needs no schedule against backgrounds bounded below its threshold, and it fails only where the background maximum reaches the threshold (Proposition 8.7). In constructed tests, the fixed softmax sink loses abstention within 0.43 nats of the predicted onset in all 4 configurations, while with weak evidence a key check holds abstention longer than the $\log N$-tempered sink (Section 8.4). With an exchangeable background, decisions on sums need a $\log N$ schedule at a fixed logit scale while decisions on maxima need none against backgrounds bounded below their threshold, and, under part (b) of Proposition 9.10, no router beats the argmax on the same scores (Sections 9.3 and 9.4). In the pilot's extension, a trained key check keeps both sides in every constructed geometry except heavy-tailed keys, where it loses both, its abstention failing at the wall its extreme-value law predicts (Section 17.2).

17.6 Recommendations and limits

Chapter 0 states the rules that can be used now (Section 0.6.1). Give attention a way to deliver nothing, a gate or a learned sink: in trained 125M language models it removes the first-token sink and most massive activations at no cost in perplexity at the training length (Section 17.4). When extending a model, keep its positions in distribution with interpolation such as YaRN. Without interpolation, a null outside the scores, such as an output gate, degrades about as fast as plain softmax, while a fixed null in the scores collapses. Scaling every logit by $\log N / \log N_{\text{train}}$ helped where positions were interpolated, in vision transformers at higher resolution and in causal language models; on top of YaRN, which already carries a log-length temperature, it hurt, and without interpolation YaRN's temperature alone hurt as well (Section 17.4 and Table 17.6). Its guarantee needs strictly separated margins (Theorem 9.6), which none of the probe's retrieval heads in the trained 125M language models met, and the pilot kept both sides over its tested range empirically rather than by certificate (Section 17.2). The rules' full treatment is left for later work.

The evidence has limits. The pilot fixes the feature geometry, and its extension varies it only among constructed geometries (Section 17.2). The production-model test was prespecified, not registered, and undecided by its rule, its slopes leaning toward compensation only descriptively (Section 17.3). And the trained language models have 125M parameters and three seeds per rule; no retrieval head reached the registered threshold, so their abstention results describe attention weights, and the 350M leg and the depth ladder are not yet measured (Section 17.4).

17.7 What holds in trained networks

The two-sided law, for a null that absent and present queries share at a fixed logit scale, and the law of sums carry their labels once, with the theorems they rest on (Section 9.7): the two-sided

Setting	Positions / comparison	Measure	Effect	Sec.
ViT-S, crops	Interpolated	Accuracy (pp)	Sink / gate with sink: 640 px: +3.5, +3.4 896 px: +4.9, +5.0 Gain in 6/6 checkpoints at each size	17.4
ViT-S, originals 896 px	Interpolated	Accuracy (pp)	Softmax: +4.87 [4.29, 5.46] Sink: +4.65 [3.94, 5.41] Gate + sink: +4.54 [3.93, 5.14]	17.4
125M LM	Interpolated $I_\lambda - I$	NLL (nats/token)	Lower by 0.20–0.50 9 of 9 checkpoints	17.4
125M LM	Interpolated Y vs. I_λ	NLL (nats/token)	Lower with YaRN in 9 of 9 checkpoints	17.4
125M LM	YaRN $Y_\lambda - Y$	NLL (nats/token)	Higher by 0.08–0.18 9 of 9 checkpoints	17.4
125M LM	Native $C - N$	Perplexity ratio D	Higher in 12 of 12 checkpoints	17.4
0.6M byte LM	Interpolated $Y - I$	Length penalty (bpb)	+0.038 to +0.075 (group means) 40/40 checkpoints worse	17.4
Exact-match pilot	No positional encoding	Null and match weights	Both kept in 8/8 heads 16–65,536 keys; no certificate	17.2

Table 17.6: Tempering has different effects across positional regimes. Each effect compares the same checkpoint with and without the stated temperature. Vision uses one ViT size on ImageNet-100, 3 seeds per rule, and interpolated positions. Crop gains are descriptive at 640 px and prespecified at 896 px; counts are checkpoints. Resized-original gains have recorded 95% paired hierarchical bootstrap intervals (seeds, then matched images); most originals are also upscaled. pp denotes percentage points. The 125M comparisons use 3 seeds per rule: $\lambda = \log N / \log N_{\text{train}}$ on interpolation (I) or YaRN (Y), and YaRN's temperature alone (C) on native positions (N). NLL ranges span checkpoints for tokens beyond training in 8,192-token windows; D is the full-window perplexity ratio to the training-length baseline. The $I_\lambda - I$ and $Y_\lambda - Y$ NLL comparisons are prespecified; the Y versus I_λ comparison and the $C - N$ count are descriptive. The byte-model range spans 8 position/rule means (5 seeds each; RoPE or partial RoPE): YaRN (Y: interpolation plus its temperature) versus interpolation alone (I), at 2,048 bytes. The length penalty uses a common native 256-byte chunk baseline; bpb means bits per byte. This comparison is descriptive; the temperature scales content scores, leaving learned nulls unscaled. The exact-match pilot uses 2 seeds in each of 4 configurations; retention is empirical, with no separated-margin certificate. That row covers the pilot's isotropic geometry: in the extension, tempering lost abstention with heavy-tailed keys at both budgets and, at the smaller budget, with clustered keys and with four 32-dimensional heads. Ranges are not confidence intervals; the measures have no common effect scale.

law is exact under premises checked network by network, and that sums need a $\log N$ schedule at a fixed logit scale while maxima need none against backgrounds bounded below their threshold is shown in an idealized setting.

Exact under premises checked network by network. No fixed rule keeps both sides at long context, for a null that absent and present queries share, and tempering the whole budget keeps both under strictly separated margins (Sections 17.1, 17.2 and 17.6; a null that may differ between the two queries is priced by Proposition 9.2): the impossibility bound is exact for any softmax row with content logits in $[-\tau, \tau]$ and a null shared by absent and present queries; tempering is guaranteed to keep both sides at every length only under strictly separated margins, which are checked on the head, and over a finite range it can keep both without them, as in the pilot.

Shown in an idealized setting. Decisions on maxima need no length schedule against backgrounds bounded below their threshold, so a head built on maxima could keep abstention and retrieval at every length (Section 17.5): shown for constructed keys and backgrounds (independent or exchangeable scores).

Measured in trained networks. The production-model sink test (Section 17.3). Prespecified on gpt-oss-20b: undecided by its prespecified rule; descriptively, its slopes lean toward compensation through position. The trained abstention test (Section 17.4). Prespecified 125M language-model test: its lead prediction (that $\log N$ tempering on top of YaRN holds abstention and lowers degradation) failed. Every rule that gives attention a null removed the first-token sink and cut massive activations, all but sigmoid lowering perplexity at the training length; positional interpolation carried most of YaRN's gain past the training length, and log-N tempering helped on interpolated positions. In its vision subset (ViT-S/16 on ImageNet-100), tempering raised accuracy at untrained token counts without restoring the trained attention. The abstention results describe attention weights, not behaviour.

Chapter 18

Bounded memory: recall costs capacity

Bounded memory is the second place where the design can still move. Chapter 0 states it in brief (Section 0.3.7): attention keeps every past token and pays for memory in full, with a cache that grows with the context (Table 9.1), and a recurrent mixer avoids that price with a state of fixed size, which buys robustness to length and pays in recall. This chapter develops the trade. Section 18.1 shows that recurrent and streaming mixers retrace the lattice of Chapter 14 with mechanisms of their own, and Section 18.2 gives the write condition that bounds their state at every length. Section 18.3 prices recall as a capacity per head, not a distance, and Section 18.4 gives the reserve that hybrids keep for it, a few attention layers whose cache stays linear in the context. Section 18.5 says why a comparison of recurrence with attention needs a qualified baseline first, and Section 18.6 reports the robustness to length measured so far, in-sample. Several of these results are left for later work, and the chapter holds their places; Section 18.7 says what holds in trained networks.

18.1 A successor class retraces the lattice

Recurrent and streaming mixers are the main alternative to attention, and they meet the same failure modes with their own mechanisms. Each protection of the inventory (Table 0.2) reappears under another name (Bruce, 2026d). A decay gate is the anchor. The delta rule is separation: it addresses by key and writes a value, and with unit keys, unit step and no decay it overwrites a key exactly and leaves orthogonal directions unchanged. A scan is ordered, so the arrow comes free. The write condition is the class's gain control (Section 18.2), and a fixed-size state pays for its length robustness in recall (Section 18.3). Hybrids keep a few attention layers as the routing-budget reserve for recall (Section 18.4), and output gates and denominator sinks carry abstention here as they do in attention (Chapter 8). Chapter 0 counts the usual answer that challengers converge towards attention (Section 0.2.3) among those the account explains: within the stated classes, successors must pay the same conditions and retrace the lattice (Table 0.1).

The decay gate acts as the anchor inside a trained hybrid (Bruce, 2026d). In a delta-rule recurrence with a short causal convolution and one output-gated attention layer in four, without the gate the log of the state's norm rises by 2.23–2.28 at base scale and 1.94–2.01 at small scale over a sequence of twice the training length (in the layer where it rises most), against 0.24–0.87 and 0.14–0.71 in the four variants that keep the gate, with no overlap across five seeds; a run without the convolution shows the same separation, although there the model without the gate did not learn the recall task at base in any of five seeds (0.178 in length). With the convolution, removing the gate costs 0.121 perplexity at base, while recall stays at 0.999 in length and falls from 1.000 to 0.940 at 64 stored pairs, beyond the maximum of 32 used in training, and no perplexity failure

appears with length. Reserving part of the state as a register costs no perplexity at base (4.318 with it against 4.344 without, at 6,613,048 parameters). The recall and perplexity values are means over five seeds. These are measurements on trained models, from the local records of that experiment, reconciled between its two runs; the environment of the cluster runs is not verified.

Latent recursion shows the anchor directly. On associative recall, a weight-tied block trained at 4 loops keeps most of its accuracy at 96 loops when its input is re-injected at every loop and loses it when it is not (three seeds). A moving average of the latent that starts at the input delays the loss (Section 5.2, Table 5.2).

Recall is expected to need an outer-product (delta) write before capacity matters, though no registered test has checked it yet, and state tracking beyond counting needs transitions that do not commute: a state updated by commuting linear maps depends only on the symbol counts (Bruce, 2026d).

Proposition 18.1 (What commuting transitions cannot distinguish)**.** *If each symbol a applies a fixed linear state transition A_a and all transitions commute, then from a common start the final state after a word depends only on the count of each symbol, and no readout of it can distinguish two words with the same counts.*

The proof is in Section A.19. With additive writes the updates are the affine matrices $\left(\begin{smallmatrix} A_a & b_a \\ 0 & 1 \end{smallmatrix}\right)$, which need not commute when the A_a do: the updates $x \mapsto 0.5x + 1$ and $x \mapsto 0.9x$ do not. Non-commutation can therefore be necessary in this restricted setting; it is not sufficient to solve every state-tracking task. The group-theoretic statement is due to Grazzi et al. (2025) and Siems et al. (2025). Chapter 0 reads the usual answer that weaknesses are patched above the architecture, with chain of thought, retrieval and tools (Section 0.2.3), as adjacent to this: state tracking beyond counting needs non-commuting transitions, and otherwise the collection is silent on it (Table 0.1).

The anchored aggregator of a frozen causal kernel streams. Its output $Y = \beta(I - \alpha P)^{-1}X$ needs no matrix inverse: row i of $(I - \alpha P)Y = \beta X$ gives each output from the earlier ones (Bruce, 2026d).

Proposition 18.2 (Streaming resolvent)**.** *Let P be a frozen lower-triangular substochastic matrix, $0 \le \alpha < 1$ and $\beta = 1 - \alpha$. The anchored output $Y = \beta(I - \alpha P)^{-1}X$ is computed in index order by*

$$y_i = \frac{\beta x_i + \alpha \sum_{j<i} p_{ij} y_j}{1 - \alpha p_{ii}},$$

and every denominator is at least $\beta > 0$. If in addition $p_{ij} = u_i^\top v_j / z_i$ for $j \le i$, with $u_i, v_i \in \mathbb{R}^r$ and known positive row normalizers z_i, then

$$y_i = \frac{\beta z_i x_i + \alpha u_i^\top H_{i-1}}{z_i - \alpha u_i^\top v_i}, \qquad H_i = H_{i-1} + v_i y_i^\top,$$

which uses $O(rd)$ persistent state for d-dimensional outputs. For normalized factorized kernels, maintaining $s_i = \sum_{j \le i} v_j$ also gives $z_i = u_i^\top s_i$ (plus any declared sink mass).

The proof is in Section A.19. One forward pass with $O(rd)$ state, independent of the sequence length, therefore gives an anchored aggregator at linear-attention cost. Constant memory means independent of the sequence length for fixed r and d; it does not apply to dense causal kernels such as softmax, whose prefix can require every earlier output. If P is rebuilt from the unknown output Y, the system is nonlinear and implicit, and the frozen argument no longer suffices. Chapter 0 lists the streaming resolvent among the results that are still partial, synthetic or unrun (Section 0.4.6): it is a design without a trained test.

18.2 Strictly contracting writes bound the state

A delta-rule write updates a matrix state M with a key k and a value v, at a step $\eta > 0$ and with a decay $\lambda \geq 0$:

$$M \mapsto M(I - \eta(\lambda I + kk^\top)) + \eta v k^\top.$$

Write M_t for the state after t writes. Vector norms are Euclidean; for matrix states, use either the spectral or the Frobenius norm. Non-expansiveness compares states written with the same key and value.

Proposition 18.3 (A common contraction factor bounds the state)**.**

(a) The write is non-expansive when $\eta(\lambda + \|k\|^2) \leq 2$.

(b) If every key has norm at most K *and every value norm at most* V*, and*

$$q = \max(|1 - \eta\lambda|,\ |1 - \eta(\lambda + K^2)|) < 1,$$

then $\|M_t\| \leq q^t \|M_0\| + \eta V K(1 - q^t)/(1 - q)$ *at every length* t.

The proof is in Section A.19. Non-expansiveness bounds only differences between states, not the state itself: without decay, unit keys whose directions converge while their values conflict can make the state grow like $\sqrt{t}$ (Section A.19). Strict contraction bounds the state at every length, and with bounded readout logits the per-token loss is bounded too; the write condition is the class's gain control (Bruce, 2026d). Without decay, $\lambda = 0$, the factor q is at least one whatever the step, so the bound needs the decay: limiting the step buys non-expansiveness, not a bounded state.

The write certificate is consistent with the observed stability outcomes in small trained fast-weight memories (Bruce, 2026d). With three seeds per configuration, none of 48 certified runs diverged, all of them on multi-query associative recall. Among uncertified runs, 41 of 45 aggressive runs diverged; of 228 tuned runs, the 12 that diverged were all language-modelling runs, a task with no certified arm, while none of the 216 tuned recall runs diverged (about 14–30M parameters overall, with training budgets of 10,000 steps for recall and 20,000 for language modelling). These are descriptive totals over different task mixes. On a synthetic memory swept across its boundary, no certified cell was classified as divergent during 50,000-token runs; the certificate is conservative for memories with momentum, and observed boundedness matched it in 97% of 270 cells when the cells with momentum that stayed bounded between the certificate and the measured stability edge count as matches. A causal test inside a trained model is still to run (Section 18.6).

The certificate has been checked for the delta-rule write and for a causal flow block built to satisfy it. Its algebra holds to machine precision on random draws: the closed form for the norm of the delta-rule transition, and the norm of the write as an operator on the state, agree to within 3.6×10^{-15}, and in 200,000 draws the step-size condition never disagreed with the computed norm about contraction or non-expansiveness. The causal flow block keeps its protections by construction, without training, in each of the nine configurations checked: its token operator is exactly lower-triangular, so no output depends on a later token, and its routing is non-expansive to machine precision. In small running delta-rule loops, the 25 cells checked used five decay settings and five seeds. Over the 400-write horizon, the contracting and marginal settings stayed bounded, while the expanding settings diverged. The confirmation in trained memories is the fast-weight test above.

What non-expansiveness bounds and what it costs make one trade: non-expansiveness bounds the differences between states but not the state, a common contraction factor bounds the state (Proposition 18.3), and a bounded state pays in recall (Section 18.3). The law that prices this trade is left for later work.

18.3 Per-head capacity limits recall, however far back

A fixed-size state pays for its length robustness in recall. A linear memory reads a value back as Mk from a key $k \in \mathbb{R}^{d_k}$.

Proposition 18.4 (Recall is a capacity per head)**.** *Every assignment of values to N keys is exactly retrievable from a d_k-dimensional linear memory if and only if the keys have rank N. So one head stores at most d_k pairs exactly.*

The proof is in Section A.19. The statement counts keys and their rank, not how long ago a pair was written: a bounded state's recall limit is a capacity, not a distance. Exact recall of N arbitrary associations therefore needs $N \leq d_k$ per head: each head's state must be sized to the associations it reads back. Chapter 0 reads the usual answer that exact, content-addressable memory makes recall possible (Section 0.2.2) as supported and sharpened: a bounded state's exact recall costs capacity per head, $N \leq d_k$, and recall is one condition of six (Table 0.1).

How efficiently a bounded state pays this price has been measured on constructed memories (Section 9.5, Remark 9.15). An outer-product state of d^2 numbers, written with random unit keys, recalls 0.315–0.328 nat per number, close to $1/\pi$, once it holds $16d$ items or more. A least-squares state, the delta rule's fixed point, recalls every stored bit exactly up to $N = d$, the capacity of Proposition 18.4, and the online delta rule with a unit update rate peaks near $N = d$ to $2d$ and then recalls what is recent.

Under one matched-data protocol, frozen before the runs, a gated delta-rule recurrence recalls 32 stored pairs at 0.998–0.999 and holds 0.997–0.999 across an 8,192-token gap (five seeds); at 128 pairs, beyond the 32 it was trained on, its recall falls to 0.78–0.83.

The recall cost of bounded state is not established. The routing-only arm of the recall-at-distance test, the minimal flow block, is bimodal across seeds with its curriculum on (with it off, both seeds stay low) and was trained at a preregistered learning rate not retuned for it (Figure 18.1), while the gated delta-rule recurrence above recalls 32 stored pairs at 0.998–0.999, within its training load (five seeds); both are measurements on trained models (Bruce, 2026d). The cost becomes a claim only after the recall-capacity sweep inside the hybrid experiment (Section 18.4).

Against the number of stored pairs, the records of the matched-data protocol hold the gated delta-rule recurrence (five seeds) and a hybrid that adds attention, in parallel with the recurrence, in two of its eight layers (two seeds), both trained on 32 pairs. At each number of pairs their recall barely changes with the filler between the stored pairs and the queries, from none to 2,048 tokens, and it falls as the pairs grow. Across those fillers both recall 32 pairs at 0.997 or above; the recurrence recalls 64 pairs at 0.969–0.985 and 128 at 0.778–0.832, and the hybrid recalls 128 at 0.809–0.852. These records hold no attention arm. Figure 18.1 shows recall against distance for every saved seed of a smaller recall test. Exact recall against the number of pairs stored, for several key dimensions of the state and with an attention reference, is left for later work.

A flat, low recall curve is not a capacity limit. Figure 18.1 keeps every saved seed of the recall-at-distance test rather than pooling a learned run with a failed one. The registered run used eight stored pairs, width 128, a frozen learning rate of 10^{-3} and two seeds per arm and curriculum. The full-flow model has 297,365 parameters and each transformer 331,905, a mixer-stack ratio of 1.1307 to the flow target. The run's own remark on learnability at a higher rate describes a reduced check (two layers, width 64, six pairs, 20 keys and 20 values, 700 steps, learning rate 3×10^{-3}), not a rerun of the registered size. The comparison at a fixed rate therefore does not establish a capacity boundary.

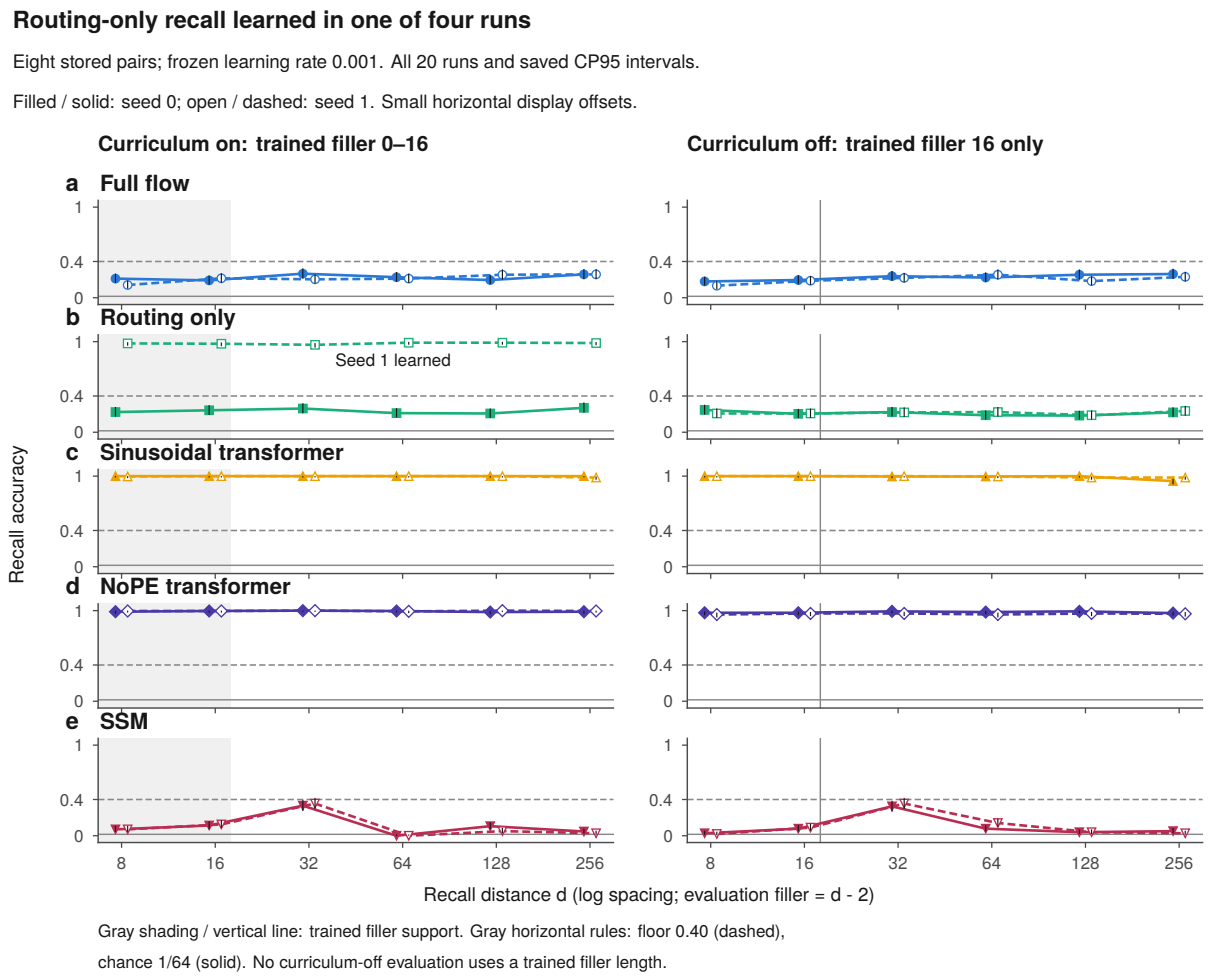


Figure 18.1: **Routing-only recall learned in one of four saved runs, while all eight transformer runs were strong.** Rows (a–e) show full flow, routing only, the sinusoidal and NoPE transformers, and the SSM; columns compare curriculum on and off. Every seed appears at all six distances: filled/solid is seed 0, open/dashed is seed 1. Small offsets of ± 0.065 in $\log_2 d$ separate coincident marks; vertical values and interval endpoints are unchanged. Thin black error bars are the saved Clopper–Pearson 95% intervals for each $k/256$ last-pair query evaluation, not uncertainty across the two model seeds; the evaluation batch at each distance is shared across cells. Gray horizontal rules mark chance 1/64 (solid) and the near-recall floor 0.40 fixed in advance (dashed). Shading on the left and the vertical line on the right mark trained filler support under evaluation's $F = d - 2$: $F = 0$–16 with curriculum, $F = 16$ without. The full trained distance envelopes are 2–35 and 18–35; sharing a distance envelope does not imply trained filler support. Strong means minimum recall across distances at least 0.80; all nine strong and eleven below-floor runs are retained, with no partial runs. Retention is uninterpretable at or below the near-recall floor and is not plotted. The 129-token argmax can yield accuracy below value-chance, including zero. These fixed-learning-rate (10^{-3}), eight-pair records (flow width and state 128) do not establish a capacity boundary or settle learnability at the registered size; the higher-rate learnability remark concerns a reduced-size check. The runtime environment was not recorded.

18.4 Hybrids keep an attention reserve

Row B2 of the lattice states the trade (Table 14.2, Section 14.1). Relaxing the unbounded key–value cache to a bounded, contracting state buys length robustness, and $O(1)$ state for a pure recurrence; what must then be re-supplied is a recall reserve, a few attention layers whose cache stays linear in the context. Non-expansive writes alone bound no state, while a decay gate, a factor $a < 1$ on the state, does for bounded keys and values (Section 18.2), and a recurrence keeps $O(1)$ state only if it has no full-context attention layers. A hybrid that keeps some stays $O(n)$, with a smaller coefficient.

Deployed hybrids keep such a reserve. Qwen3-Next keeps full attention in 12 of its 48 layers, and Kimi Linear in 7 of its 27 (Qwen Team, 2025; Kimi Team, 2025), a descriptive record read from the models' published sources (Section 16.3). Whether such fractions sit at corners is a hypothesis, and its reading of the hybrid sweep below is pending (Section 16.4).

The state and the cache can be counted exactly. From 1,024 to 8,192 tokens of context, the attention arm of the matched-data protocol keeps a key–value cache that grows in proportion to the context, while the gated delta rule keeps no cache, only a state and a short convolution buffer of fixed size. A hybrid with attention in two of its eight layers keeps a quarter of the attention arm's cache, still linear in the context, beside the same state; at 8,192 tokens its cache and state together are 3.93 times smaller than the attention arm's cache. Wall time is another matter. The delta mixer runs as the sequential reference recurrence, so every measured cell of an arm that contains one, 18 of the 21 measured cells, is flagged as inadmissible for a speed claim; only the attention arm's timings are admissible, and a comparison of wall time waits for a verified chunkwise kernel. Chapter 0 finds the collection silent on the usual answer that training is parallel, with every position giving a signal (Section 0.2.2), beyond this: contracting recurrences also parallelize, and wall-time claims need verified kernels (Table 0.1).

Whether a few attention layers buy recall back, and at what price in quality at the training length and in degradation at eight times it, is a frontier against the attention fraction. Tracing it at two scales, with exact recall beside quality, is left for later work.

Pending: The qualified hybrid sweep at 125M and 350M parameters: a hybrid of a certified contracting recurrence ($a < 1$) with gated attention under the registered positional treatment, swept over its attention fraction, with quality at parity, degradation at eight times the training length, state size, the certificate's contribution and the recall-capacity sweep that the recall cost of bounded state waits for; Section 16.4 reads it against the corner hypothesis.

18.5 Qualify the baseline

Section 18.3 reported the gated delta-rule recurrence of the matched-data protocol alone. That protocol's attention baseline is not yet qualified, so no ranking between the two is claimed (Bruce, 2026d).

An attention baseline that has not been qualified voids the wins recorded against it: one recorded win of recurrence over attention rested on an attention baseline that converged to a key-blind prior, scoring 0.078–0.081 recall at 32 stored pairs and 128 filler tokens in four of five seeds (recovered from the logs of runs that did not finish), where a key-blind predictor was estimated to score 0.085; its full treatment is left for later work. Small-scale wins can therefore fail validity checks, a second reason, beside the compensation that Chapter 0 gives (Section 0.2.1, Table 0.1), to read the usual answer that proving an alternative at scale costs too much as refined.

The protocol behind these comparisons matches the data. Attention, the gated delta rule and hybrids that keep attention in two of their eight layers are trained on one task, multi-query

associative recall, for the same number of steps with the same batch, width and depth, and they are scored on one grid of stored pairs and filler lengths. A full statement of its validity gates, the correctness suites, numerical thresholds by depth and precision, and analytic checks that a run must pass before it is compared, is left for later work.

The protocol's records also test overwriting and interleaving at 32 stored pairs, with 128 filler tokens for the unique-key and overwrite evaluations and none for the interleaved queries. Trained on a mix of examples with unique keys and examples that rewrite a quarter of their stored keys and ask for the latest value, the gated delta rule and two hybrids have recall on examples with overwrites within 0.003 of their recall on unique-key examples in every seed, with unique-key recall of 0.998 or above; the attention arm, trained on the same mix, stays at 0.080–0.085 on unique keys in all five seeds. Trained instead on records and queries that interleave, every arm answers the interleaved queries at 0.96 or above, the attention arm included, while recall of unique keys after that training ranges from 0.02 to 0.98 across arms and seeds. Transfer without retraining, from unique keys to these two distributions, is not recorded.

Attention may also serve recurrence during training rather than at inference, as a scaffold: a teacher whose outputs or hidden states supervise a recurrent student and which is removed at evaluation. The records hold one seed with a teacher, and that teacher is the attention arm of the protocol above, whose baseline is not yet qualified. Supervised by it, gated delta-rule students recalled 64 pairs with 512 filler tokens, beyond the training maximum of 32 pairs, at 0.944–0.972, against 0.985–0.991 without a teacher. Both ranges span two student variants; the runs without a teacher span two seeds. Whether a qualified teacher helps is left for later work.

Pending: Byte-level language modelling at matched data: bits per byte on WikiText-103 at 1,024 to 8,192 tokens of context for six arms, on a pilot budget and then on the main one, with their parameter counts.

18.6 In-sample length robustness

The existing evidence bears on length, not on recall. A certified non-expansive flow block degrades 1.24–1.56× in bits per byte at eight times its training context, against 2.18–4.22× for transformers with sinusoidal or no positional encoding and 49–75× for a diagonal state-space model, the exploding arm of the comparison, whose transition has $|a| = 1$ (Bruce, 2026d). These are in-sample measurements on trained models, with loss averaged over every position of that context and ratios averaged over two seeds at each of two scales, against weak baselines, sinusoidal and no-position transformers only, and the largest model has 19.2M parameters.

Figure 18.2 shows every saved seed. Both flow arms have lower in-sample loss than both tested transformers at two, four and eight times the training context in every saved seed at both scales, about 3.29M and 19.2M parameters, while the sinusoidal transformer is better at the training length; the diagonal state-space model degrades about 49 and 75 times at eight times. The text is WikiText-103, scored in-sample on the training slice and averaged over all positions, and these cross-architecture curves establish neither held-out generalization, nor a causal certificate effect, nor a bounded-state result.

Pending: Whether the certificate is causal inside a trained model: degradation at eight times the training context and growth of the state's norm against the write strength s, with and without decay, including the no-decay non-expansiveness boundary at $s = 2$.

Pending: How the gap between the flow block and the transformer at the training length changes with scale.

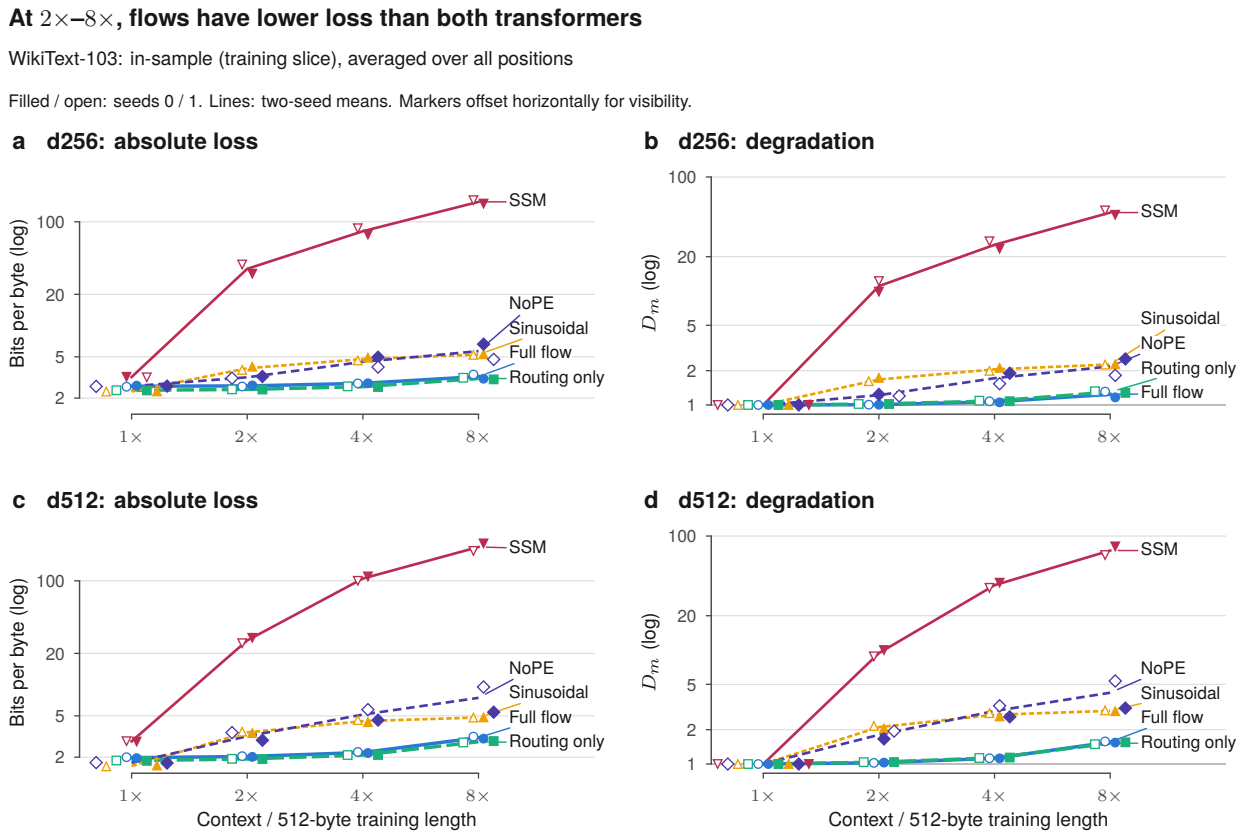


Figure 18.2: **Both flow arms have lower in-sample loss than both tested transformers at 2×, 4× and 8× context in every saved seed at both scales; the sinusoidal transformer is better at 1×.** WikiText-103, in-sample (training slice), averaged over all positions. (a,c) Bits per byte; (b,d) degradation D_m, relative to the same seed at 1×. Both vertical axes are logarithmic. Rows are flow-target d256/L12 and d512/L18; total parameters are approximately 3.29M at d256, and 19,173,722 (full flow) against 19,198,208 (transformers) at d512. Full flow, routing-only flow, sinusoidal and no-position (NoPE) transformers, and a diagonal selective state-space model (SSM) are shown. Filled and open markers show seeds 0 and 1, offset only horizontally to prevent overlap; lines show saved means. There are 20 cells (5 arms × 2 scales × 2 seeds), with 32,768 predicted bytes per context. Windows are shared across arms and seeds at each context but differ across contexts; no intervals or p-values are reported. The d512 full-flow 1× mean is 1.98 (seeds 1.957 and 1.998). D_m averages seed ratios, not means before division. Where the two conventions differ by more than 0.005, mean-of-ratios against ratio-of-means are: d256 SSM: 2× 11.0366/11.0297, 4× 25.4094/25.3983, 8× 48.7695/48.7570; d512 SSM: 4× 37.0560/37.0485, 8× 74.6275/74.6023; d512 NoPE: 8× 4.2218/4.2278. The original runtime environment was not recorded. These cross-architecture curves do not establish held-out generalization, a causal certificate effect or a bounded-state result.

18.7 What holds in trained networks

Qualified baselines carry their label once, with the leak rule, in the closing of the chapter on the arrow (Section 10.9): measured on trained baselines, with the leak rule itself exact.

Exact under premises checked network by network. Bounded memory, which buys robustness to length and pays in recall (Sections 18.1, 18.2, 18.4 and 18.6): exact for linear memories (through the rank of the keys) and for delta-rule writes with bounded values and decay under the stated step-size condition, for any sequence of keys bounded in norm. Recall as a capacity per head, not a distance (Section 18.3): exact for linear read-back from a d_k-dimensional memory (through the rank of the keys) and, for the state bound, for writes that contract by a common factor.

Chapter 19

Beyond the transformer: machines and the price list

Chapters 17 and 18 looked for room where Chapter 0 places it: where a condition is patched rather than met, or paid at a price that another design could avoid (Section 0.3.7). This chapter moves from the transformer's room to the design space. The conditions of Part II were read from the transformer, but they describe a design space, not a design. The chapter states them without the transformer (Section 19.1), prices machines that pay them in other currencies, physical devices among them (Sections 19.2 and 19.3), reports where small trained machines sit relative to their walls, with their other trained tests still to come (Section 19.4), lists what remains of the room inside the transformer (Section 19.5) and ends with what any successor must supply (Section 19.6). Section 19.7 says what of this holds in trained networks.

Chapter 0 states the price list and the catalogue of machines in brief (Section 0.3.3), and what a successor must supply (Section 0.3.7). This chapter develops them from the catalogue of routing machines and the lattice of protections (Bruce, 2026e,d), without repeating that summary.

19.1 Seven conditions that do not assume a transformer

The six protections that Part II proves and Chapter 14 organizes into a lattice are each a conditional statement within a declared class of models, on the operator theory of Part I: what goes wrong, and at what price, when a model in that class lacks the property (Bruce, 2026d). None of them needs softmax, matrices or layers. Stated without those words, they become conditions that any machine must meet in which content decides who talks to whom, something else decides what is said, and the step repeats (Bruce, 2026e). A seventh condition joins them: transport alone cannot create features, so every machine needs a writer (Section 11.1).

Each condition is a test. Removing it leaves a signature that a measurement can see, and a machine can pay for it in more than one currency. Every proposition that prices a payment is proved within a declared class, and this book states each with its premises. The measurements in Sections 5.4, 6.3, 7.6, 8.4, 9.3 and 10.7 are constructed tests that check the propositions, not trained-model results, except where a section says otherwise. Two results organize the rest. The law of sums (Propositions 8.7 and 9.11) says that a decision summing over all N items at a fixed logit scale needs a $\ln N$ schedule against exchangeable backgrounds, while a decision that reads a maximum needs none against backgrounds bounded below its threshold. In the Gaussian surrogate, a sum whose scale grows past the point where the background sum is carried by its largest terms reads the maximum, at the price of a scale that grows with N (Section 9.4). The price of order

Table 19.1: The conditions as tests. These six are the protections of the lattice of Chapter 14, with the names that Chapter 0's inventory gives them in parentheses where they differ (Section 0.3.2); the writer, the seventh condition, is stated as a test in Section 11.1. A machine meets a condition by paying in any of its currencies, and the result named in the last column prices the payment.

Condition	Removal signature	What to measure	Currencies	Priced by
A tether (the anchor)	Soft routing reaches consensus; hard routing coalesces onto its cycles	Identity retention; spread; distinct states	Restart probability, residual share, leak conductance, retained endowment	Proposition 5.1
A routing budget	Dilution (soft) or misses (hard) once $\ln N$ outgrows it	The failure onset $\ln N^*$ against the currency	Logit gain; $\ln M$ samples with a select readout; tree depth; address bits; and the query's evidence, which no router replaces	Propositions 9.10 to 9.12
A null (abstention)	Forced answers: absent queries retrieve something	False retrieval on absent queries against N	A competitive null logit; an absolute threshold; an empty bucket; an outside option	Proposition 8.7
Separate address and cargo (alignment)	Contamination: routing scores pick up payload terms	The onset shift, tied against untied	Channel width; factorized addresses	Theorem 6.11
Gain below the tether (gain control)	Runaway: several fixed points, high sensitivity, lock-in	Contraction, multiplicity, sensitivity	Logit scale times squared radius, relative to the anchor; comparator sharpness	Proposition 7.7
An arrow	Order blindness: soft routers fail every latest-write query, hard ones half	Latest-write accuracy, and its price on old items	Positional range; masks; timestamps; overwrite; driven cycles	Proposition 10.6

(Proposition 10.6) says that a hard machine orders writes without spending any of the router's margin, while a soft one pays in positional range (Section 10.7).

Table 19.1 states the six protections as tests: the signature that each one's removal leaves, what to measure, the currencies in which a machine can pay, and the result that prices the payment (Bruce, 2026e). The writer's row, the seventh, is stated as a test in Section 11.1.

Read this way, the lattice is a design space: a machine meets each condition by paying in any of its currencies (Bruce, 2026e). Figure 9.7, in Section 9.4, draws three of the choices on exact constructions. A sum pays for the size of its population while a maximum does not; a head that routes over its top k keys and gates on its largest score separates routing from abstention, with no explicit length schedule built in; and order carried in soft weights has a positional price, while hard timestamps or overwrite order every pair without spending positional-logit margin. These are illustrative constructions, not experimental measurements; in trained 125M language models with YaRN, such a head was the only tested rule whose null and match weights both changed by less than half a nat of log-odds out to eight times the training length, on mean per-seed changes over 3 seeds though not in every seed, and it failed natively (Section 17.4). These are attention weights from fallback heads, not behavioural abstention or retrieval.

19.2 Machines pay in different currencies

The currencies are already in hand. Samples with a select readout, and the hard stage of a tree, buy a soft router gain (Proposition 9.11, Section 9.3). Address bits and key width pay the routing budget (Table 19.1; Section 9.3). State pays the memory ledger; timestamps pay for order in storage,

while overwrite gives up the history instead (Table 9.1, Section 9.5). On a physical substrate, the dynamic range of a device caps what one soft stage can spend (Proposition 19.1, Section 19.3). Within its class, each wall moves only when its own ledger is paid (Corollary 9.14). The ledgers are those of the conservation of routing information, a family of lower bounds: each holds within its own class of machines and prices one currency there, and the bounds set no exchange rate between currencies (Section 9.5).

A catalogue places 43 families of routing machines on the conditions, in ten groups by the currency in which each pays the routing budget (Bruce, 2026e). Soft routers pay it in logit scale, sampling machines in samples through a select readout, and hard, addressed and indexed machines on the query's evidence or in address bits. No family buys evidence, so every machine that ranks scores stops at the evidence wall at the latest (Section 9.3). A family inside a proposition's class is predicted by it; for the others the mapping is a reading, and the physical, biological and social systems in the catalogue are hypotheses or analogies. This is the catalogue to which Chapter 0 points for how other machines pay for the same conditions (Section 0.3.3).

19.3 Other substrates

Two conditions can also be read off physical substrates: the routing budget, through a device's dynamic range, and the tether, through a passive circuit (Bruce, 2026e). These are statements about ideal devices, not measurements: convergence rates need specified capacitances and dynamics, and no device was simulated. Read onto a real physical, chemical or biological machine, each is a hypothesis.

The first statement prices the routing budget in a device's dynamic range.

Proposition 19.1 (A dynamic-range wall for one competitive stage)**.** *Let a stage normalize N positive conductances whose ratio of largest to smallest is at most R. Its target's weight is at most $R/(R+N-1)$, so a target weight of at least $w \in (0,1)$ needs $N \le 1 + R(1-w)/w$. The bound is sharp, so at fixed target weight and large R a stage buys about $\ln R$ nats of $\ln N$.*

Proof. Let $g_{\min}$ be the smallest of the conductances. The target's conductance is at most $Rg_{\min}$ and the other $N-1$ are at least $g_{\min}$, and the target's share is largest at these extremes. Solving the resulting weight inequality gives the count. □

A shunt, the circuit's null, adds competing conductance and cannot raise the target's weight. Hard zeros, pruning and independent stages change the class: the bound applies to one stage, not to a whole cascaded device. It is the bound of part (c) of Proposition 9.12, with the conductance ratio in the place of the exponentiated logit window, which is $e^{2\tau}$ for cosine scores at logit scale τ. So the on/off ratio of an ideal device caps the gain a single soft stage can spend.

The second statement finds the tether in a passive circuit.

Proposition 19.2 (A passive reciprocal circuit is a tethered router)**.** *Let $W = W^\top \ge 0$ hold reciprocal edge conductances, with zero diagonal and positive degrees $d_i = \sum_j W_{ij}$, and let $D = \operatorname{diag}(d_i)$. Connect node i to a clamped source at potential x_{0i} through a conductance $g_i = \beta d_i/\alpha$, with $0 < \alpha < 1$ and $\beta = 1-\alpha$. Kirchhoff's steady-state equations have the unique solution $x = \beta(I-\alpha P)^{-1}x_0$ with $P = D^{-1}W$, which is the tethered fixed point of Proposition 5.1. For arbitrary positive shunts g_i, the solution is $[I - \operatorname{diag}(\alpha_i)P]^{-1}\operatorname{diag}(1-\alpha_i)x_0$ with $\alpha_i = d_i/(d_i+g_i)$.*

Proof. Current balance at node i reads $\sum_j W_{ij}(x_i - x_j) + g_i(x_i - x_{0i}) = 0$. With proportional shunts this is $[(D-W) + (\beta/\alpha)D]x = (\beta/\alpha)Dx_0$, and multiplying by αD^{-1} gives the stated resolvent,

whose inverse exists by the convergent Neumann series of αP, whose rows sum to $\alpha < 1$. For general shunts, dividing row i by $d_i + g_i$ gives the second form; its substochastic matrix has largest row sum $\max_i \alpha_i < 1$, which gives uniqueness in the same way. □

Reciprocal passive conductances make P reversible, with stationary weights proportional to the degrees d_i. A directed, nonreversible routing kernel needs further circuit elements. So, with proportional shunts, an ideal relaxation network computes the tethered router's fixed point exactly, and Proposition 5.1 applies to it unchanged; an arrow, if it needs one, must come from elements that break reciprocity (Section 10.1). Analog crossbars already solve such linear systems, PageRank among them, in one step per right-hand side, though with operational amplifiers in feedback rather than a passive network (Sun et al., 2019), and nonlinear resistive networks have been trained end to end with equilibrium propagation in circuit simulation (Kendall et al., 2020).

For a stochastic physical machine, the arrow also has an energetic price. If its routing is a Markov kernel with a positive stationary law, circulation costs entropy at second order in the current, whatever its sign (Theorem 10.8, Section 10.7). That is a statement about a physical machine's Markov dynamics, not a dissipation claim for attention: an attention matrix by itself is not a measured dissipation.

19.4 Small trained tests and machines at scale

Small trained tests. Seven routing machines were trained end to end as the token mixer of otherwise identical small pre-normalized models, beside softmax, on CIFAR-10 at 32 pixels and on byte-level WikiText-2 with a 128-byte context, in models of about 141 thousand and 594 thousand parameters (Bruce, 2026e). Every machine trains, and ten of the twelve informative pairs land within the registered parity band of softmax, so at this scale trainability does not single out softmax. Length transfer differs by machine. Evaluated on 512 bytes, the three machines with an explicit length rule (a decay, a $-\ln n$ offset and an in-layer restart) show no penalty, while the other four all pay one. The walker pays most, as a fixed sample budget suggests (Proposition 9.11), and letting its samples grow with the context removes most of that loss. A log-N logit scale learned in training does not remove the penalty of the competition machines it was trained with and raises it for softmax and the tree, so the registered structural reading is refuted, and the transfer is reported machine by machine. Against a sliding window with no resets, every machine except the bounded state does worse with the full context. From 32 to 48 pixels every image mixer loses 3.4 to 5.7 points of accuracy.
Pending: The routing machines at scale.

Stage A: the small trained models. A battery of readouts has read three walls off the small trained models of the machines (Bruce, 2026e): the tether's coalescence (Section 5.4), the gain certificate (Section 7.6) and the evidence wall of the routing budget (Section 9.3). Its first stage, Stage A, covers 135 cells, 100 of them language models and 35 image models, with every gate passing (Table 19.2). By its registered rule the stage is inconclusive. No head qualified as a retrieval head in any language-model cell (a qualifying head needed a mean weight of at least 0.3 on the target at the training context, a threshold set before scoring), so the two retrieval predictions have no decided head; that is a qualification finding, and it licenses no conclusion about evidence walls or gain. With each routing matrix frozen and iterated without a tether, hard-routed image mixers, the tree, hash-bucket (LSH) and walker mixers with 5 paired seeds each, lose identity while spread stays high in 0.000 of heads, against 0.000 for softmax: there is no trained support for coalescence missed by spread, though the coalescence of Section 5.4, measured on constructed keys, is not contradicted. The median trained head has $\hat{\alpha}(1 + 2\tau R^2)$, the basic certificate of Proposition 7.7(a) with a measured $\hat{\alpha}$ in the place of α, between 13.3 and 35.2 across the 14 applicable groups (the

walker counted once for each of its two sampling rules; per-seed medians 12.4–39.2), against the prediction's band of 0.1, set well below the certificate's bound of 1, with measured $\hat{\alpha}$ medians of 0.13–0.16 in the language models and 0.33–0.38 on images: trained heads sit far outside the certified region in general, not only at retrieval heads. The certificate is sufficient, not necessary (Section 7.6). The battery's later stages are not reported here.

Table 19.2: Stage A of the readout battery on the small trained models of the machines: 135 cells (100 language-model, 35 image), every gate passing. The registered verdict is inconclusive: the two retrieval predictions have no decided head. κ is the certificate's quantity $\hat{\alpha}(1 + 2\tau R^2)$ with the measured update share; the certificate needs $\kappa < 1$, and the prediction's band is 0.1.

Registered prediction	Groups	Stage A	What decides it
Softmax retrieval fails at the evidence wall	Softmax language models, plain and with scalable softmax	inconclusive	no seed contributes a decided head (0 qualified retrieval heads)
The law of Proposition 9.10(c) predicts the argmax onset	As above	inconclusive	as above
Hard-routed image mixers coalesce where spread misses it	Tree, LSH and walker; 5 paired seeds	fails (all three)	spread-miss rate 0.000 against 0.000 for softmax; 95% interval [0, 0]
Trained τR^2 far below the certificate	14 groups (5 image, 9 language-model)	fails (14 of 14)	no seed has median $\kappa \leq 0.1$ (4 of 5 image or 8 of 10 language-model seeds required); per-seed medians 12.4–39.2

19.5 The room inside the transformer

This section keeps to the room inside the transformer: where one of its own conditions is patched rather than met, or paid at a price that another design could avoid (Section 0.3.7). Abstention and retrieval at long context are Chapter 17's. Machines beyond the transformer, its recurrent successor class included, are those of Chapter 18 and Sections 19.1 to 19.4.

Pending: Direction as a knob: learned per-head circulation gates, whether the learned gate's direction agrees with the first-order predictor, and how predictable such knobs are, for organized blocks and for circulation gates.

Direction from composing reversible layers is the subject of Section 10.6: two heads reversible under one common law circulate in composition exactly when they do not commute, and whether such a composition routes is the test that section awaits.

In a decoder the reversal objects are anti-causal. For a causal softmax head that gives every earlier token positive weight, the stationary law sits on the first token, and a reversal, reversible part or circulation built from any regularized positive law mixes later tokens into earlier ones (Bruce, 2026d). The cautionary case is a GPT-2 with its backbone frozen and rank-64 flow-and-resolvent layers trained for three epochs on WikiText-2. It reached validation perplexity 18.09, against 30.08 before training, but a separately trained arm with its reversible part and circulation causally re-masked reached only 23.43, and the resolvent alone 23.50: the extra gain over the resolvent was a leak from future tokens (one seed). The rule this case motivates is the decoders' rule of Section 10.8: never apply J, S or π in a decoder without a lower-triangular mask and a leak test that perturbs the tokens after position i and requires zero change at i.

Blocks whose preference, exchange, circulation, abstention and anchor are explicit parameters have their construction identities checked to machine precision before training, roughly match a parameter-matched transformer at the training length on three small synthetic causal tasks, and,

in-sample on WikiText-103, degrade less at eight times their training context than transformers with sinusoidal or no positional encoding (Section 18.6). The identities are checked before training; the other two are measured on trained models.

The completion and replication of the minimal anchored block against tuned baselines, and which of its factors carries the gain, are the tests that Section 16.7 awaits.

Stability by construction, coupling the logit scale to the anchor, against QK-norm, logit soft-capping, z-loss and σReparam, is the test that Section 7.5 awaits.

What did not work is printed, for the most part, where each protection is: pure transport without writing (Section 11.1); engineered null registers, and eviction of cache entries by absorbed mass, judged in a linear model built from captured attention (Section 8.5); and compact value-channel designs, which saved cache but gained neither task performance nor decoding speed (Section 14.1). Free signed normalization on homophilous graphs did not work either, measured on trained models (Bruce, 2026d); signed mixing itself is the subject of Section 15.5.

Three more negatives complete the list (Bruce, 2026d). Per-instance measured anchor dials were an honest tie once priced. Two further negatives rest on weaker records. Coupling damping with rotation in selective recurrences won none of four memory tasks (five seeds; a win required beating every structural ablation on average), a number from a laboratory log whose stores were never synced. And in an unregistered comparison on CIFAR-10 (eight layers, width 128, three seeds), the free value map added 3.99 percentage points of accuracy to attention, while swapping attention for the resolvent, with the value map removed from both, changed it by -0.06 points. The lattice anticipates most of these: pure transport cannot write; abstention needs only a way to reduce the update, not a particular sink; signing buys subtraction, which homophilous graphs do not need; the value map carries the gain; and circulation rotates without amplifying. Eviction by absorbed mass is a limit of the tested policy, which the lattice does not address.

19.6 What any successor must supply

Within the stated classes the inventory is a checklist. A successor keeps short-path mass, a free content path, gain control of the full feedback Jacobian, a way to abstain whose budget survives length, and a routing budget that grows like $\log N$; it takes its arrow from circulation, a mask or composition; and it must write. It will win, if it wins, mainly on a resource axis at quality parity, and the lattice says which protection it must re-supply for each relaxation it makes (Section 14.2) (Bruce, 2026d). Machines that meet the conditions without attention are those of the catalogue of routing machines, whose currencies Sections 19.2 and 19.3 give (Bruce, 2026e). Chapter 0 reads two of the usual answers through this checklist (Section 0.2.3). Challengers converge towards attention: within the stated classes, a successor must pay the same conditions, and so retraces the lattice, as the recurrent class does (Section 18.1). The transformer absorbs its challengers as components: most of what gets absorbed into the mixing layer re-supplies a named protection, which is a reading (Section 16.5; Table 0.1).

What decides among the machines is the price list. The transformer is the cheapest point under one price list, not the only one, and the sweet spot is a judgment under that list, not a computation (Section 0.3.3; Sections 14.4 and 14.5). Underneath the list sits the conservation of routing information: a family of lower bounds, each holding within its own class of machines and pricing one currency there, with no exchange rate between currencies (Section 9.5). How the list decides among the machines of this chapter is left for later work; Bruce (2026e) reads it under assumed prices.

A successor must write, and what makes a good writer is open; the main open problem the

framework leaves is how much of each protection's price the MLP can absorb, and at what width (Section 11.7). Where each routing machine of the catalogue makes its features, and which writer it pays with, is still to be reported there. Section 21.3 lists the writer with the other open problems.

19.7 What holds in trained networks

Three claims that this chapter draws on carry their labels where they are developed: the conservation of routing information in Section 9.7, the price list in Section 14.6, and qualified baselines with the leak rule for decoders in Section 10.9.

An interpretation, with nothing to transfer. The checklist for a successor, the seven conditions read without the transformer (Sections 19.1 and 19.6): a checklist read from the stated classes; there is nothing to transfer.

Chapter 20

Predictions stated in advance

Parts IV and V offer four readings under a model, which are not statements about networks: the price list and the sweet spot (Sections 14.4 and 14.5), the adopted changes read retrospectively (Section 16.5) and the checklist for a successor (Section 19.6). There is nothing for them to transfer, and none of the adopted changes counts as evidence, since none was predicted before it was adopted. An interpretation stands or falls with predictions stated before the evidence (Section 0.5). Chapter 0 lists the predictions in brief (Section 0.6.2); this chapter gives them in full. Section 20.1 states the ten predictions, the labels fixed in advance for the trained-harness outcome, and the reported result. Section 20.2 gives the prediction tested on a production model and what each of its outcomes would have changed. Section 20.3 summarizes the corner-fraction scoring convention, and Section 20.4 registers every decisive test still open, with what each outcome would change.

20.1 The predictions and their tests

The predictions below come from the architecture capstone (Bruce, 2026d). Three have been tested: the sink length law, on a production model, where the test was undecided by its prespecified rule (Section 20.2), and in the trained 125M harness, where the fixed sinks took the registered compensated branch; sigmoid retrieval, in the same harness; and, provisionally, transport-only mixers. The other seven have not been tested and are not yet evidence for the account (Section 0.3.8); a reader can hold each against the models that follow. Each is falsifiable with public models or small controlled runs. Ten predictions follow, in the order of Chapter 0's brief list, and then the labels fixed in advance for the tempered rule's outcome in the trained harness, with that outcome.

Sink length law. In dense layers of models with learned sink logits, the log-odds of sink mass fall with a slope near -1 in the logarithm of the query position, and layers with a banded window are flat beyond their window. The test on gpt-oss was undecided by its prespecified rule: its books control failed, and its slopes matched neither prespecified branch (Sections 17.3 and 20.2). Because the sink's log-odds equal b minus the background's log-sum (Section 17.3), their slope is $-\gamma$ by construction, where γ is the growth of that log-sum, so the prediction is about γ itself: near 1 without positions, as in small byte-level models above the shortest tested training length; below 1 inside the training length with rotary positions, as in trained 125M language models, where a learned sink's median head slope, averaged over three seeds, is -0.12 on FineWeb-Edu with native positions (Sections 9.2 and 17.4); and lower in more local heads. A rotary head whose background grows at the full rate inside its training length, or a more local head whose background grows faster than a global one's, would count against it. That form is fitted to those models, so only other models can test it.

Norm control accompanies dropped row sums. Any widely adopted attention that drops unit row sums ships with a norm-control mechanism. In the lattice, relaxing unit row sums must re-supply norm control when normalization is dropped (row A1 of Table 14.1; Section 14.2). Current designs already pair the two, for example retention with group normalization (Sun et al., 2023) and sigmoid attention with hybrid normalization (Ramapuram et al., 2025); the prediction is for future designs, and a widely adopted attention that drops unit row sums with neither output nor logit normalization would refute it.

Sigmoid retrieval. Sigmoid attention with a $-\log n$ bias loses single-key retrieval weight like $1/N$ unless match logits grow like $\log N$; with content logits in $[-\tau, \tau]$, fixed positive abstention and match weight cannot both hold beyond $N \propto e^{2\tau}$ (Bruce, 2026c). Both statements rest on Proposition 9.3, which assumes one offset shared by the matched absent and present queries. In the trained 125M harness with YaRN, sigmoid match weight fell from the training length to eight times it in all three seeds, by 1.14 in log weight on average against the prespecified minimum drop of 1.0, though one seed fell by only 0.48; a strict $1/N$ law would give a drop of $\ln 8 \approx 2.08$. No head met the retrieval-identification threshold, so this concerns attention weights. The same registered prediction expected the calibrated sink to lose its match weight; with YaRN it lost its null mass instead (Section 17.4).

Signed gains shrink with heads. Gains from signed or differential attention shrink as the number of heads grows at fixed width. Differences of heads already realize subtraction inside multi-head attention, so a signed head adds capacity mainly when heads are scarce (Section 15.5). The test is pending.

Circulation in diffusion language models. Masked diffusion language models carry more delivered circulation in early layers than masked-language-model encoders of similar size. Their attention is bidirectional, so it is eligible for the measures of circulation, whereas a causal decoder's stationary circulation vanishes (Section 3.3); delivered circulation is that of Section 10.2, and Section 10.3 measures circulation in two masked-language-model encoders. The test is pending.

Orientation needs odd statistics. A statistic that is invariant under stationary reversal cannot see the sign of a reversal's effect; for smooth readouts of one amplified operator this is a theorem (Theorem 3.4). The prediction concerns statistics not yet proposed: any future statistic that predicts the sign of reversing a trained head's circulation, the reversal of Section 10.2, will change under reversal.

Logit-scale feedback. Self-referential logit instabilities are predicted where the feedback outweighs the anchor, beyond the certificate $2\alpha s R^2 < \beta$ for tied dot-product feedback on a domain of radius R (Sections 7.1 and 7.2). If the training update is modelled with $\beta = \eta\lambda$ and $\alpha = \eta$, the certificate's threshold is $s_c = \lambda/(2R^2)$: at a fixed radius it scales with the weight decay alone, and the learning rate cancels (Section 7.3). Because the certificate is only sufficient, this is a weak test: an onset of logit growth that moves with the learning rate at fixed weight decay would count against it. The test, the coupling against the standard fixes, is pending (Section 7.5).

Corner fractions. Sequence models keep the fraction of their layers that use full attention at 0, at 1, or near a small-denominator level in $\{1/2, 1/3, 1/4, 1/6, 1/8\}$, scored by the distance to the nearest level against a null (Section 16.3); the level set is a convention, not a consequence of the theory. On a single attention-fraction axis, concave pricing predicts 0, 1 or a binding cap, never a fraction such as $1/4$ in itself, so a small fraction counts as a corner only by this convention (Section 16.2). That designs jump rather than drift is a hypothesis: the deployed fractions do not yet discriminate between corners and drift (Sections 16.3 and 16.4). Chapter 0 calls this the list's weakest test (Section 0.6.2).

Anchor horizon in looped models. For a fixed operator with $0 < \mu_C < 1$ and a fixed per-visit mixing share p, a mode's gain falls tenfold after about $2.3/(p(1-\mu_C))$ visits when $p(1-\mu_C)$

is small (Section 5.1); the prediction is that this horizon describes content loss in trained loops without re-injection. For M distinct tied blocks each visited N times, in the fixed-operator model of Proposition 2.14, a block whose update has symmetric fraction λ_s multiplies its leading-order energy loss per visit by $1 + 2(N-1)\lambda_s^2$. The update scale $c/(N\sqrt{M})$ suffices for every symmetric fraction; $c/\sqrt{MN}$ suffices when $\lambda_s \quad N^{-1/2}$. For a fixed operator and a mode $0 < \mu < 1$, restart re-injection with fixed weight r keeps the floor $r/(1-(1-r)\mu)$ at every loop count, with no horizon, whereas accumulating the input without weights, which is uniform averaging once normalized and when each visit mixes the whole accumulated state, still forgets, at rate $1/((L+1)(1-\mu))$ (Sections 2.1 and 2.2); that law need not carry over to a Pre-LN residual loop. The shares p, λ_s and r are measured, not fitted. The test is pending.

Transport cannot write. Mixers that only transport, preserving each token group's channel Gram matrix, are predicted to underperform on retrieval-and-write tasks at any depth. In the tested family of conditional orthogonal transports, a writing channel restored CIFAR-100 classification but not dictionary retrieval, a provisional comparison because implementation qualification and matched-parameter controls remain unresolved. The statement concerns pure transport; variants that re-center or write break its invariant (Remark 11.2, Section 11.4).

In the trained harness of Section 17.4, the outcome for the $\log N$-tempered rule was to be labelled with one of four branches fixed in its protocol before the run: certified (the separated margins of Theorem 9.6 hold on the measured scores), explained by background concentration, non-stationary (the score law drifts with length), or empirical over a stated length range. All four presupposed that tempering kept both sides. At 125M, tempered on top of YaRN, the rule did not keep abstention and raised long-context perplexity, so the outcome lies outside all four, and no head was certified. In the controlled pilot the tempered rule kept both sides over the tested range empirically, not by certificate (Section 17.2).

20.2 The sink length law on a production model

The test of the sink length law on a production model is one of the decisive tests that have run. On gpt-oss the law became a prespecified, not registered test of three legs, which Section 17.3 prints with its decision rule and its result. All three legs holding would confirm the law; the dense leg failing while the two banded legs hold would mean that the model compensates; anything else is inconclusive. The recorded outcome is undecided: the control that must pass before any leg is read failed on the books, and the slopes fell outside both decisive branches, though the dense layers' falling background leans toward compensation (Section 17.3).

Each outcome would have changed the account differently. A confirmation would have given the account of abstention at long context (Chapter 17; Section 0.3.7 in Chapter 0) a production-model confirmation, stated before the run. Compensation would have meant that background logits fall with length: the law holds, but trained models adapt. Because the outcome is undecided, evidence for the law in trained heads comes from the controlled pilot of Section 17.2 and the trained models of Section 17.4. The trained harness has since reported at 125M (Section 17.4): on text, fixed sinks compensated through position, inside the training length and with YaRN beyond it, but not natively, and the tested small byte-level models with a learned null and no positional encoding follow the stationary law within each model, except those trained at the shortest length.

20.3 The corner predictions

The corner predictions come from the corner hypothesis of Section 16.2. For attention fractions they score a release against 0, 1 and the fractions $1/K$ for $K \in \{2, 3, 4, 6, 8\}$, the level set of the corner-fraction prediction (Section 20.1). That level set is the scoring convention of Section 16.2, not a consequence of the theory.

20.4 The register of open tests

Table 20.1 is the register of the decisive tests. It is generated from the claims ledger of Appendix B, which also gives Chapter 0's tables, and for each test it gives what the test measures, its status, what each outcome would change in the account of Chapter 0, and the chapters that hold its place. Two tests have returned: the prespecified test of a production model's sinks, which was undecided by its rule, its slopes leaning toward compensation through position only descriptively (Sections 17.3 and 20.2), and the trained language-model test of abstention rules at 125M (Section 17.4), where the $\log N$ tempering prediction failed on top of YaRN, while the registered prediction that an output gate, alone or with a learned sink, lowers the first-token sink and massive activations held at 125M (it failed in the vision subset, and the 350M leg has not run). The 350M leg and the other four tests await their main results. The paragraphs below take them in the order of the chapters that hold their places.

Can any capacity replace a removed anchor? With the identity path removed, MLP width, heads and depth are swept against arms that re-supply an anchor another way. The outcome tests the anchor exception of Section 5.5 in trained networks: that capacity elsewhere does not buy back short-path mass at gain bounded independently of depth under strict mixing, although large gain at finite depth can recover content (Section 0.3.4 in Chapter 0).

Does compensation grow with scale? The capacity a tied value map costs is measured in a ViT-S and in a 125M-parameter language model. The outcome tests the scale hypothesis of Section 15.6, the proposed mechanism for why the small-scale gains of alternatives fade.

Does the minimal block's margin survive tuned baselines? The 22M-parameter comparison of Section 16.7 is replicated with a clean implementation against baselines trained with the DeiT recipe. If the margin survives, the blueprint (Section 16.6) gains a data point at scale. If it shrinks to parity, it is stated as parity, and nothing the architecture capstone claims changes, since it claims no win over tuned transformers at scale (Bruce, 2026d).

Which abstention rules keep both sides in trained language models? Language models of 125M parameters were trained with each abstention rule, the max-statistic head included. The learned sink and the gate with a learned sink were also tempered at evaluation, and abstention and retrieval were measured out to long context (Section 17.4). At this scale tempering on top of YaRN, which already carries a log-length temperature, is not a measured fix: it raised long-context degradation, did not hold abstention, and no head had the separated margins its guarantee needs. With YaRN the max-statistic head, which reads maxima rather than sums (Section 9.4), was the only rule whose null and match weights both changed by less than half a nat of log-odds out to eight times the training length, on means over 3 seeds though not in every seed, and natively it failed as the matching key fell far down the ranking and the background maximum grew with length, the evidence wall of Section 9.3. The 350M leg is not yet measured.

The qualified hybrid ratio sweep. The fraction of attention layers in a certified recurrent hybrid is swept at 125M and 350M parameters against qualified baselines (Sections 18.4 and 18.5). The outcome decides whether bounded memory with an attention reserve is a win on a resource axis

Table 20.1: Decisive tests: what each measures, its status, what each outcome would change in the account of Chapter 0, whose sections the fourth column cites, and the chapters that develop it. Generated from the claims ledger (Appendix B).

Test	What it measures	Status	What each outcome would change	Chapter
Do a production model's learned sinks stop abstaining with length?	Measures whether gpt-oss-20b's sink log-odds fall with context length in dense layers and stay flat in banded ones beyond their window.	*measured*: undecided by its prespecified rule (its books control failed, and its slopes matched neither branch), though descriptively its slopes lean toward compensation through position.	Confirm: the account of abstention at long context (Section 0.3.7) gains a production-model confirmation, stated before the run. Compensated: background logits fall with length; the law holds but trained models adapt. Inconclusive: the controlled pilot and the trained harness carry the evidence.	17 and 20
Which abstention rules keep both sides in trained language models?	Trains 125M language models with each abstention rule, including the max-statistic head, and measures abstention and retrieval out to long context. The learned sink and the gate with a learned sink are also tempered at evaluation.	*measured*: the lead prediction failed, as tempering on top of YaRN, which already carries a log-length temperature, raised long-context degradation and did not hold abstention; and the max-statistic head was the only rule to keep both weights within half a nat on the seed means with YaRN, and failed natively.	Decides whether tempering is stated as a measured fix at scale, and whether the max-statistic head escapes the log-N length schedule (Sections 0.3.7 and 0.4.4).	17 and 20
What fraction of attention layers does a qualified recurrent hybrid need?	Sweeps the fraction of attention layers in a recurrent hybrid with certified contracting writes at 125M and 350M against qualified baselines that pass the recall check.	*pending*	Decides whether bounded memory with an attention reserve is a resource-axis win (such as memory or speed) at quality parity (Section 0.3.7).	18 and 20
Does the minimal block's accuracy margin survive tuned baselines?	Replicates the 22M-parameter comparison with a clean implementation against DeiT-recipe baselines.	*pending*	Survives: the blueprint (Section 0.3.6) gains an at-scale data point. Shrinks to parity: stated as parity; no win over tuned transformers is claimed. Loses: the 22M result is reported as not replicated, and the blueprint rests on the small-scale record.	16 and 20
Does compensation grow with scale?	Measures the capacity a tied value map costs against MLP ratio, width and depth in vision transformers on ImageNet-100 and a 125M language model.	*pending*	Tests the proposed mechanism for why alternatives' small-scale improvements fade (Section 0.3.4).	15 and 20
Can any capacity replace a removed anchor?	Removes the identity path and sweeps MLP width, heads and depth against arms that re-supply an anchor another way.	*pending*	Tests the anchor exception in trained networks (Section 0.3.4).	5 and 20

at quality parity. Read against the corner hypothesis, the same sweep is a planned test of that hypothesis (Section 16.4).

20.5 What holds in trained networks

Pending: there is nothing yet to transfer. The untested predictions stated in advance (Sections 20.1 and 20.3). Predictions about trained networks, stated before their tests; the tests already run are reported with their outcomes.

Closing

Chapter 21

What this changes

Chapter 0 ends by saying what the account changes for architecture work, and where it stops (Section 0.6): a way to read a proposal, the rules that can be used now, the predictions stated in advance and the limits, each in brief. Chapter 20 gives the predictions in full. This chapter gathers the rest from the chapters that develop it: the rules (Section 21.1), the limits (Section 21.2), what would change the account and the problems it leaves open (Section 21.3), and what the collection does not address (Section 21.4). It ends where Chapter 0 ends (Section 21.5).

21.1 Rules usable now

Table 21.1 collects the closed-form design rules that follow from the results of Chapter 2 and Section 8.3 (Bruce, 2026k). The objects for which those results are exact, and beyond which they are hypotheses, are stated in Section 2.9.

The operator theory of Part I has a practical reading of its own (Bruce, 2026i). Deep averaging needs short-path mass (Theorem 2.8); residual stacks with a constant update share only slow collapse (Proposition 2.20). On the lazy, head-averaged operators measured in Section 3.6, the reversible part of one attention operator is a close proxy for its collapse-type quantities, not for its eigenvectors, and it is no proxy for a stack of distinct layers. The order of amplification and a second map decides which of two channels is open (Theorem 4.1); on a lazy reversible operator an anchor caps every gain ratio and how far the amplifier can turn the seed (Corollary 4.2).

Each rule that Chapter 0 lists as usable now (Section 0.6.1) has its chapter, which gives its premises and its evidence: giving attention a null, and keeping a model's positions in distribution when it is extended, in Section 17.4, where scaling every logit by $\log N / \log N_{\text{train}}$ is also measured (its premises in Section 9.2 and its controlled pilot in Section 17.2), with the rules stated as recommendations in the first paragraph of Section 17.6; bounding a recurrent state, with bounded keys and values, by decay that makes every write contract by a common factor rather than by limiting the step, and sizing each head's state to at most d_k arbitrary associations it must read back exactly from a linear memory, in Propositions 18.3 and 18.4 (Sections 18.2 and 18.3); scaling coherent residual branches by $1/L$ and incoherent ones by $1/\sqrt{L}$, with L the total depth, in Theorem 2.13 (Section 2.4); masking and testing for leaks any decoder layer built from a reversal or a flow in Section 10.8; and reading test-time context mixing with an affine readout through each prediction's own and companion margins in Section 12.5.

Table 21.1: **Design rules.** Closed forms for keeping a mode μ at gain at least γ, or for keeping a floor as depth grows; the example column uses $\mu = 0.85$, $\gamma = \frac{1}{2}$, $L = 64$.

Decision	Rule	Example
Restart / input injection r	$r \geq \gamma(1-\mu)/(1-\gamma\mu)$, valid at every depth	$r \geq 0.1304$
Lazy residual weight p at depth L	$p \leq (1-\gamma^{1/L})/(1-\mu)$	$p \leq 0.0718$
Diffusion time t (heat, DGC)	heat: $t \leq \log(1/\gamma)/(1-\mu)$; L Euler steps: $t \leq L(1-\gamma^{1/L})/(1-\mu)$ with $L \geq t$	heat: $t \leq 4.621$; Euler: $t \leq 4.596$
Layer averaging	gain $\sim 1/((L+1)(1-\mu))$; add a fixed input atom a for a floor a	gain 0.103
Residual branch, centered weights	$\alpha_L = c/\sqrt{L}$: floor $e^{-c^2\sigma^2(1-\mu^2)}$	—
Residual branch with drift m	need $\sup_L L(\alpha_L m + \alpha_L^2\sigma^2) < \infty$, i.e. $\alpha_L = O(1/L)$	—
Tied loop, M blocks $\times$ N visits	$\varepsilon = c/(N\sqrt{M})$ unless the symmetric fraction $\lambda \quad N^{-1/2}$	—
Suppress μ_- below η, keep μ_+	at most $w\mu_+^k + (1-w)\mu_+^{k+1}$ (Proposition 2.22); otherwise go signed	0.4932 for $(0.5, 0.9, 0.01)$
Rollout diagnostics	measure $p_{\ell,i}$; report the sink treatment; survival $= \|B_L \cdots B_1 \mathbf{1}\|_\infty$	—

21.2 Limits

Chapter 0 states the account's limits in brief (Section 0.6.3); this section gathers them from the chapters, beginning with scale. The largest controlled test is the trained harness at 125M parameters, whose 350M leg is not yet run, and much of the controlled evidence is below 25M (Bruce, 2026d). The production-scale evidence is measurement on public checkpoints: captured operators at short context, sixteen pretrained language models (Section 2.7), and the gpt-oss test, which was prespecified, not registered, and undecided by its rule, its slopes leaning toward compensation through position only descriptively (Section 17.3). What compensation implies for a protection's price measured at small scale is the hypothesis of Section 15.6, whose test is pending.

That the standard block is where every protection is cheap rests on a semi-formal price model, not on measured prices (Section 14.4); the first paragraph of Section 14.5 says why the sweet spot is therefore an interpretation under that model, not a measured optimum.

Reading past design changes through the conditions is retrodiction, not evidence: none of the changes was predicted before it was adopted (the fourth paragraph of Section 16.5), and only the predictions stated in advance can test the reading (Chapter 20).

The reversible vocabulary needs a positive stationary law, which full causal softmax attention over more than one token lacks; its direction comes from the mask. Section 10.8 lists the statements proved for decoders and the leak rule for layers built with a reversal, J or S from a regularized positive reference law. Nor is there a theory here of what the MLP should compute, beyond that it writes and that it is the main currency of compensation (the second paragraph of Section 11.7).

One limit bounds the account as a whole. Chapter 0 states it (Section 0.6.3) and returns to it in its closing paragraph, which also closes this book (Section 21.5): the conditions bind layers that take weighted averages, with weights rebuilt from their input and stacked through depth, at today's prices for hardware and training; a design outside that class, or different prices, could favour an architecture that looks nothing like a transformer; and the account explains why the transformer has lasted, not that it will.

Three readings in the operator theory are interpretations rather than results (Bruce, 2026i): the link between the structural collapse target and the position where trained models place attention sinks, which the theorem does not explain (Section 2.5); the reading of the GPT-2 composite as a function of an assumed share (Section 2.7); and the rule for choosing the order of amplification and a second map from what the map reads (Section 4.3). Which kernel constructions enter the harmful placement regime is left open (Section 4.3). Explicit abstention mechanisms, such as sink registers or softmax variants whose rows may sum to less than one, are not treated in that theory; Chapter 8 takes them up.

Each source paper states its own scope, and the book prints each statement with the closing of the chapter that develops the paper: for the operator theory (Bruce, 2026i), in Sections 1.6, 2.9, 3.7 and 4.6; for depth as a distribution (Bruce, 2026k), in Section 2.9; for converged averages (Bruce, 2026h), in Section 7.8; and for circulation (Bruce, 2026g), in Section 10.9.

An interpretation, with nothing to transfer. That most controlled evidence is small, and that the evidence at production scale is measurement on public checkpoints (the first paragraph of this section; Section 0.6.3): a statement of scope. That the account says nothing outside its class or at other prices (Section 0.6.3): a statement of scope.

21.3 Open problems and what would change the account

In the operator theory of Part I, what can fail is the theorems' reach (Bruce, 2026i). That theory would be weakened by (i) trained causal models in which vanishing sink weight is common, which would make the envelope of Theorem 2.16 vacuous where it matters; and (ii) attention operators on which collapse-type readouts move as much, in relative terms, as orientation-type readouts at native flow, or on which the first-order formula of Theorem 3.4(iii) fails to predict the eigenvector turn. A separate open test is whether a real operator with $R > 1$ after deflation shows amplify-then-project doing worse than project-then-amplify, the contamination signature of channel (iii) of Theorem 4.1. Real operators with $R > 1$ exist, but those measured did not show a clear signature (Section 4.3).

The operator theory leaves open problems. What law governs the collapse of a trained causal stack, with attention rebuilt at every layer and MLPs and normalization in between? What is the effective update share of a trained residual stream? In Pre-LN transformers the stream grows with depth while each block's update does not, so deep blocks contribute less (Sun et al., 2025). A share that falls with depth does not fix the regime by itself: in the mean-field initialization model, an incoherent stream has an energy share falling like $1/\ell$ and an amplitude share falling like $\ell^{-1/2}$ at once and forgets only polynomially, while a stretched-exponential loss, $\exp(-\Theta(\sqrt{L}))$, needs that share profile together with uniform mixing and a coherent drift (Section 2.6; Bruce, 2026k). Which regime trained stacks occupy is a measurement, and the exact consensus/contrast log of Section 2.9 can make it. Which kernel constructions enter the harmful placement regime, and what does placement do there? A signed filter pays for escaping the characterization in coefficient variation, which, for a fixed suppressed band and a fixed separated higher band kept near unit gain, must grow at least, and need grow at most, polynomially in the suppression achieved (Proposition 2.9); what are the sharp exponent and constants of that price? And does the separation of the two readouts at native flow, measured on lazy, head-averaged operators from two small encoders (Section 3.6), hold across model families and scales?

The writer remains open: what makes a good writer, and how much of each protection's price the MLP can absorb, and at what width, are stated as open in the second paragraph of Section 11.7; the checklist for a successor, which must write, points back there (Section 19.6).

Table 21.2 lists every result still to come in this edition, chapter by chapter, in the words of the

paragraph that holds its place; the one result checked numerically rather than proved, the stationary power-mean law, is in Section 17.2. The decisive tests, and what each outcome would change in the account, are the register of Section 20.4.

Table 21.2: **What is still to come in this edition.** Every paragraph of the book that is held for a result not yet in, chapter by chapter, in the words of its Pending line. One result is checked numerically rather than proved: the stationary power-mean law of Section 17.2. The decisive tests, with what each outcome would change, are the register of Section 20.4.

Section	Still to come
Chapter 4	
4.4	The headroom flip on a second dataset and backbone: propagation gain against label density on neighbour graphs of CIFAR-10 and ImageNet-100 features, with the deflated advantage and the cells the gauge admits.
4.5	How graph-consistency regularization turns a response to perturbation into class structure: the neighbour graph is the operative factor and the perturbation secondary, and the gain depends on cluster structure, headroom and balance.
Chapter 5	
5.5	Whether any capacity can replace a removed anchor: with the identity path removed, MLP width, heads and depth are swept in trained networks against arms that re-supply an anchor another way.
Chapter 6	
6.5	Learned and discovered blocks in trained models: cross-segment influence and accuracy against depth and the number of segments, for dense, masked, sliding-window, local, NSA-style and learned-block attention.
6.5	Document masks on real tasks: accuracy, or exact match and F1, and the measured cross-document leak of each attention variant on multi-document question answering, long-document classification and packed pretraining.
6.5	How much pretrained language models leak across documents: cross-document attention mass per head and composed influence across depth, in public language models on packed sequences, with and without the sink columns.
Chapter 7	
7.5	Coupling the logit scale to the anchor, enforced by a spectral cap on the query–key map, against the standard fixes at 125M parameters: maximum stable learning rate, learning-rate sensitivity, instability incidence and final loss without a fix, with the coupling, and with QK-norm, logit soft-capping, z-loss, σReparam, and QK-norm with z-loss.
7.7	The mechanism, tested by an experiment that separates feedback from the anchor's relaxation: where instability sets in, in the plane of logit scale and weight decay, for a softmax Gram loop, against the certified boundaries ($2sR^2$ and $8sR^2$) and the boundary set by the measured gain, with a frozen-kernel control and a control that attains the bound.
7.7	Where logit growth destabilizes language-model training: instability incidence against learning rate and weight decay at proxy scale and at 125M parameters, with the contour that the certificate predicts.
Chapter 10	
10.6	Whether two reversible mixers composed without an intervening nonlinearity route a cyclic successor relation with a linear readout, which a single reversible head cannot route (Section 10.1), and whether the composition needs a shared cyclic coordinate.
10.6	The outcome of that test, with the qualifier on whether a shared coordinate is necessary read separately for each arm.

Section	Still to come
Chapter 11	
11.6	Dictionary-retrieval R^2 and CIFAR-100 top-1 for established reversible and orthogonal designs, measured under the same protocol as the arms of Chapter 11.
11.7	Where each routing machine of the catalogue makes its features, and which writer it pays with.
Chapter 12	
12.5	Independently selected adaptive and frozen policies, compared on the full image test set, on graphs over confirmation seeds and on memory streams.
12.7	When mixing helps: the gain from propagation and the rate at which companions overturn correct predictions, against label agreement set at random, under frozen and under adaptive mixing.
12.7	Repairs for the focal failure, each fixed in advance: accuracy and resolvent self-weight against mixing strength for each repair, in the focal setting and in ordinary settings.
12.8	Patches, text and trained attention: signed class-margin influence maps and boundary F1 under frozen, adaptive and repaired patch mixing.
12.8	How often companions overturn correct predictions in practice: flip accounting across context-mixing methods, with the net gain, the harmful and helpful flips, and the share of harmful flips that companions overturn.
Chapter 13	
13.4	Root solvers can recover repelling roots that none of the sampled starts reaches within the tested budget: root-solver success, residual and the attraction of the returned root, the spectral radius of the update's Jacobian there, by solver and system class.
13.5	The response of a K-step computation: the finite-horizon response norm and the directional gain against the budget K, with the equilibrium response and the outer spectral radius, for the duplicate-pair example of Section 13.5 and for a bank of systems.
13.9	Trained looped transformers and implicit models: the settling residual, the resolved response, the finite-budget derivative against K and the attraction of the root, with and without input re-injection, and per-model rates of under-resolved responses, start-dependent endpoints, non-attracting roots and finite-budget peaks, with intervals.
13.9	A collapse observable for recursive training: lineage growth and validation perplexity against the retained fraction of real data, and the separation of candidate collapse observables against a frozen five-sigma gate.
13.9	The pencil in a trained two-loop system: collapse incidence against the measured pencil margin, the two per-loop margins and the scalar composite margin.
Chapter 15	
15.3	Which currency repairs a pruned model: recovery after head pruning, under three pruning criteria at a quarter and at half of the heads, when only the LayerNorm and readout, the MLP, attention, low-rank adapters on the MLP or on attention, or the whole model is fine-tuned.
15.4	Can a sign-flexible filter shed the anchor? The flat-band case: suppression against constant-mode conditioning for absolutely monotone and sign-flexible graph filters, with the price bound and the region of an anchor-free escape.
15.4	Whether the sink-anchored decoder's trade survives at larger scale: its gain in perplexity at 2,048 tokens against its cost at the training length, at about 77M and at 350M parameters, beside a tuned length-robust baseline.
15.5	Head count at scale: the perplexity gain of signed mixing over softmax, and the learned signed share, against head count at 125M and 350M parameters.
15.6	The registered test of compensation at scale: the tax of tying the value map against MLP ratio, width and depth, in vision transformers on ImageNet-100 and in a 125M-parameter language model, against the budget measured at small scale.

Section	Still to come
Chapter 16	
16.4	The controlled comparison of jump against drift: a learned attention share against its price, under linear and under convex pricing, with a control whose optimum is planted in the interior.
16.4	The corner reading of the qualified hybrid sweep of Section 18.4: quality, robustness at length and state size against the attention fraction, and the fraction that optimizes quality at the training length against the price of state.
16.7	Replication against tuned baselines: the minimal block against ViT-S and DeiT-S trained with the DeiT recipe at matched parameters and epochs, with a separate ResNet-50 reference that is untuned and not matched in parameter count.
16.7	Which factor carries the gain: accuracy changes under selected one-factor moves from the block toward a tuned ViT-S and from the tuned ViT-S toward the block, with the anchor varied in strength rather than deleted.
16.7	Resolution transfer for the minimal block and the ViT, with and without the $\log N$ budget: top-1 accuracy against evaluation resolution, and at 22 million parameters after the fix.
Chapter 17	
17.4	What dropping sum-to-one costs at depth: loss-spike incidence and the per-block Jacobian norm against depth, at 12, 24 and 48 layers, for softmax, off-by-one, output-gated and sigmoid attention, with and without a hybrid norm.
17.4	Where trained machines sit relative to their walls, the evidence and gain walls of Section 9.3 among them, beyond the first readout on small trained models (Section 19.4).
Chapter 18	
18.4	The qualified hybrid sweep at 125M and 350M parameters: a hybrid of a certified contracting recurrence ($a < 1$) with gated attention under the registered positional treatment, swept over its attention fraction, with quality at parity, degradation at eight times the training length, state size, the certificate's contribution and the recall-capacity sweep that the recall cost of bounded state waits for; Section 16.4 reads it against the corner hypothesis.
18.5	Byte-level language modelling at matched data: bits per byte on WikiText-103 at 1,024 to 8,192 tokens of context for six arms, on a pilot budget and then on the main one, with their parameter counts.
18.6	Whether the certificate is causal inside a trained model: degradation at eight times the training context and growth of the state's norm against the write strength s, with and without decay, including the no-decay non-expansiveness boundary at $s = 2$.
18.6	How the gap between the flow block and the transformer at the training length changes with scale.
Chapter 19	
19.4	The routing machines at scale.
19.5	Direction as a knob: learned per-head circulation gates, whether the learned gate's direction agrees with the first-order predictor, and how predictable such knobs are, for organized blocks and for circulation gates.

21.4 What the collection does not address

Chapter 0 acknowledges once what the collection does not address (Section 0.2.4): scaling laws and their economics, data, capital and hardware prices, in-context learning beyond what single heads and short compositions do, chain of thought, and what the MLP computes. Three of the usual answers are ones on which the collection is silent (Sections 0.2.1 and 0.2.3; Table 0.1): that loss falls as a smooth power law in model size, data and compute, which made capability predictable

and purchasable, on which it has no results; that in-context learning turned one model into a platform; and that training is parallel, on which it notes only that contracting recurrences also parallelize (Section 18.4). A fourth, that weaknesses are patched above the architecture, with chain of thought, retrieval and tools, it touches only through state tracking beyond counting, which needs non-commuting transitions (Section 18.1).

21.5 The close

Chapter 0 closes on components and roots (Section 0.6.4): most protections in use were found to be needed when a component broke, and asking what each fix protected, and keeping what failed along the way, is how the collection began to learn what holds the transformer in place. The chapters have developed, case by case, what that question led to, and it is not repeated here. Chapter 0's last paragraph is printed again below as the book's last words, unchanged except that its two references now point to Chapter 20 and to the register of Section 20.4; it is the book's one intended repetition.

This project's ambition has outlasted my endurance. An account this broad is surely wrong in places, partly because it is this broad, so I would ask readers to poke and prod at it; the predictions stated in advance (Chapter 20) and the tests still pending (Section 20.4) are the quickest ways to find where. Part of why the transformer abides is fast matrix multiplication, engineering and history; part, I have argued, is that it pays cheaply for what attention needs. That explains its past, not its future. Transformers will keep changing where room remains, and an architecture that needs none of this, or buys it more cheaply, need not look anything like one.

Appendix A

Proofs by chapter

Proofs follow their sources. A proof that a source gives with its statement stays with that statement in the chapters. The proofs that a source gives in an appendix are collected here, chapter by chapter and, within each chapter, in the order of the statements, together with the few supporting statements that a chapter points to without printing them. A proof that serves statements in two chapters is given once, under the first of them, and every proof is headed by a pointer back to its statement. The conventions and a toolbox of standard facts that the proofs use come first (Section A.1).

A.1 Conventions and tools

The proofs use the facts on this list, each standard and each either proved here in a line or cited. The contour-integral bound on spectral projectors (Proposition A.4) is proved outside the list; it is used only to interpret the measurements of Section 3.6.

Conventions. $D_\pi = \operatorname{diag}(\pi)$ with $\pi > 0$; $\langle u, v\rangle_\pi = \sum_i \pi_i \bar{u}_i v_i$; $M^\dagger = D_\pi^{-1} \bar{M}^\top D_\pi$, which is $D_\pi^{-1} M^\top D_\pi$ for real M; M is π-self-adjoint if $M^\dagger = M$ and π-antisymmetric if $M^\dagger = -M$. $\|M\|_\pi = \sup\{\|Mx\|_\pi : \|x\|_\pi = 1\}$ (the supremum is the same over real and complex x for real M); $\|A\|_\infty = \max_i \sum_j |A_{ij}|$ and $\|A\|_1 = \max_j \sum_i |A_{ij}|$ are the ℓ_∞ and ℓ_1 operator norms, and $\|v\|_1 = \sum_i |v_i|$ for a vector; $\|W\|_2$ is the Euclidean operator norm; $\|A\|_F$ the Frobenius norm. $\operatorname{spec}(A)$ is the spectrum, $\rho(A)$ the spectral radius, and $x \perp_\pi y$ means $\langle x, y\rangle_\pi = 0$.

T1. *Spectral calculus.* A π-self-adjoint M has real eigenvalues and a π-orthonormal eigenbasis $\{v_k\}$ (conjugate by $D_\pi^{1/2}$ to a symmetric matrix and apply the spectral theorem). For a power series g with $\sum_k |c_k| \rho(M)^k < \infty$: $g(M) v_k = g(\mu_k) v_k$; $g(M)$ is π-self-adjoint when the c_k are real; $\|M\|_\pi = \max_k |\mu_k|$, and the same holds for M restricted to the span of any subset of eigenvectors; and $\{\langle x, Mx\rangle_\pi : \|x\|_\pi = 1\}$ is exactly $[\mu_{\min}, \mu_{\max}]$.

T2. *Cauchy–Schwarz.* $|\langle u, v\rangle_\pi| \le \|u\|_\pi \|v\|_\pi$, hence $|\langle u, Mv\rangle_\pi| \le \|u\|_\pi \|M\|_\pi \|v\|_\pi$.

T3. *Operator norms.* $\|A\|_\infty$ equals $\sup\{\|Ax\|_\infty : \|x\|_\infty \le 1\}$ over real or complex x; $\|\cdot\|_\infty$ and $\|\cdot\|_\pi$ are submultiplicative; and $\|M^\dagger\|_\pi = \|M\|_\pi$, because $\|M\|_\pi = \sup\{|\langle u, Mv\rangle_\pi| : \|u\|_\pi = \|v\|_\pi = 1\}$ and the adjoint swaps the slots. Adjoints reverse products: $(MN)^\dagger = N^\dagger M^\dagger$.

T4. *Geometric series and resolvent identity.* If $\|A\| < 1$ in a submultiplicative norm, $(I - A)^{-1} = \sum_k A^k$ with $\|(I-A)^{-1}\| \le (1 - \|A\|)^{-1}$; and $(I-A)^{-1} - (I-B)^{-1} = (I-A)^{-1}(A-B)(I-B)^{-1}$.

T5. *Stationary distributions.* Every row-stochastic P has a stationary distribution: any limit point of the averages $\frac{1}{n}\sum_{k<n} u^\top P^k$ of a probability vector u is one. If P is irreducible, a nonzero nonnegative stationary vector μ is strictly positive: if $\mu_j = 0$ then $0 = \mu_j = \sum_i \mu_i P_{ij}$ forces $\mu_i = 0$ for every i that steps to j, and irreducibility spreads the zero to every token. The stationary distribution is then unique. If μ is a signed stationary vector, then $|\mu|^\top P \geq |\mu^\top P| = |\mu|^\top$ entrywise and both sides have the same total, so $|\mu|$ is stationary, and so are $\mu_\pm = \frac{1}{2}(|\mu| \pm \mu)$. For the difference μ of two stationary distributions, μ_+ and μ_- have equal mass and disjoint supports, so neither can be strictly positive unless both vanish; hence $\mu = 0$ (Levin & Peres, 2017).

T6. *Jensen's inequality.* For concave φ, $\mathbb{E}\varphi(X) \leq \varphi(\mathbb{E}X)$, with equality for strictly concave φ only if X is almost surely constant.

T7. *Triangular matrices.* The eigenvalues of a triangular matrix are its diagonal entries; products of lower-triangular matrices are lower triangular, with diagonal equal to the product of the diagonals; and $\left(\begin{smallmatrix} 1 & 0 \\ r & Q \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & 0 \\ r' & Q' \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 0 \\ r+Qr' & QQ' \end{smallmatrix}\right)$.

T8. *Total variation.* For probability vectors, $d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\sum_j |\mu_j - \nu_j| = 1 - \sum_j \min(\mu_j, \nu_j) = \max_A(\mu(A) - \nu(A))$, and d_{TV} is convex in each argument under mixtures.

T9. *Dobrushin contraction.* For row-stochastic Q with uniform leak $\lambda(Q)$ (Definition 2.17) and probability rows μ, μ': $d_{\mathrm{TV}}(\mu Q, \mu' Q) \leq (1-\lambda(Q)) d_{\mathrm{TV}}(\mu, \mu')$; for any real vector v, $\operatorname{diam}(Qv) \leq (1-\lambda(Q))\operatorname{diam}(v)$ with $\operatorname{diam}(v) = \max_{i,i'} |v_i - v_{i'}|$; and every eigenvalue $\zeta \neq 1$ of Q has $|\zeta| \leq 1-\lambda(Q)$ (Doeblin, 1937; Dobrushin, 1956; Seneta, 1981). (For the first, write $\mu - \mu' = s_+ - s_-$ with disjoint supports and common mass d; then $(\mu-\mu')Q = d^{-1}\sum_{i,i'} s_+(i)s_-(i')(Q_{i\cdot} - Q_{i'\cdot})$ and each row difference has total variation at most $1-\lambda(Q)$. For the last, let $Qv = \zeta v$ with v non-constant (constant vectors have eigenvalue 1). For complex x, $\max_{i,i'}|x_i - x_{i'}| = \max_\phi \operatorname{diam}(\mathrm{Re}(e^{-i\phi}x))$, and Q commutes with taking $\mathrm{Re}(e^{-i\phi}\cdot)$, so the real bound gives $|\zeta| \max_{i,i'}|v_i - v_{i'}| \leq (1-\lambda(Q)) \max_{i,i'}|v_i - v_{i'}|$.)

T10. *Rank.* $\operatorname{rank}(AB) \leq \min\{\operatorname{rank} A, \operatorname{rank} B\}$.

T11. *Skew matrices.* If J is real and π-antisymmetric, $\langle x, Jx\rangle_\pi$ is purely imaginary for every complex x, $-iJ$ is π-self-adjoint in the complex sense, and therefore $\{\langle x, Jx\rangle_\pi : \|x\|_\pi = 1\} = i[-\|J\|_\pi, \|J\|_\pi]$; for real x, $\langle x, Jx\rangle_\pi = 0$.

T12. *Taylor's theorem.* For twice differentiable F, $F(t) = F(0) + tF'(0) + \frac{1}{2}t^2F''(\xi)$ for some ξ between 0 and t; for matrix-valued F, $\|F(t) - F(0) - tF'(0)\| \leq \frac{1}{2}t^2 \sup_{|s|\leq|t|}\|F''(s)\|$.

T13. *Simple eigenvalues.* If A is π-self-adjoint with a simple eigenvalue μ_j and unit eigenvector v_j, then for small $|t|$ the matrix $A + tB$ has an eigenvalue $\mu_j(t)$ and eigenvector $v_j(t)$, analytic in t, with $\mu_j'(0) = \langle v_j, Bv_j\rangle_\pi$ and, normalizing $\langle v_j, v_j(t)\rangle_\pi = 1$, $v_j'(0) = \sum_{l\neq j} \frac{\langle v_l, Bv_j\rangle_\pi}{\mu_j - \mu_l} v_l$ (Kato, 1995). (Differentiate $(A+tB)v_j(t) = \mu_j(t)v_j(t)$ at $t=0$ and take inner products with v_j and v_l.)

T14. *Moment bound.* For $t \geq 0$, $\Pr[X \geq m] \leq e^{-tm}\mathbb{E}[e^{tX}]$.

T15. *Accretivity and the logarithmic norm.* If $\mathrm{Re}\langle x, Ax\rangle_\pi \leq \omega\|x\|_\pi^2$ on an A-invariant subspace, then $\|e^{sA}\|_\pi \leq e^{s\omega}$ there for $s \geq 0$ (differentiate $\|e^{sA}x\|_\pi^2$).

T16. *Cumulants.* If $\mathbb{E}K^3 < \infty$, then $\ln \mathbb{E}[e^{-sK}] = -s\,\mathbb{E}K + \frac{1}{2}s^2 \operatorname{Var} K + O(s^3)$ as $s \to 0$ (Taylor-expand the cumulant generating function).

T17. *Poisson approximation.* If $X \sim \mathrm{Bin}(n,p)$ and $Y \sim \mathrm{Poisson}(np)$, then $d_{\mathrm{TV}}(X,Y) \le np^2$ (Le Cam, 1960); and the laws Poisson(a) and Poisson(b) are within $|a-b|$ in total variation.

T18. *Bhatia–Davis and Popoviciu inequalities.* A random variable with values in $[m,M]$ satisfies $\operatorname{Var} X \le (M-\mathbb{E}X)(\mathbb{E}X-m) \le (M-m)^2/4$, since $\mathbb{E}[(M-X)(X-m)] \ge 0$ rearranges to the first inequality (Bhatia & Davis, 2000).

T19. *Gelfand's formula and Perron's eigenvalue.* $\|A^L\|^{1/L} \to \rho(A)$ in any norm; and if P is irreducible, 1 is an algebraically simple eigenvalue of P (Seneta, 1981).

A.2 Proofs for Chapter 1

Proof of Proposition 1.1. $\exp(s) = \sum_{k\ge 0} s^{\circ k}/k!$, and every Hadamard power of a positive-semidefinite matrix is positive semidefinite by the Schur product theorem ($k=0$ gives the all-ones matrix). P is similar to $D^{-1/2}KD^{-1/2}$, which is positive semidefinite, so its eigenvalues are real and nonnegative; they are at most one because P is stochastic. Detailed balance holds with $\pi \propto D\mathbf{1}$. □

Proof of Lemma 1.2. (i) $\langle x,\, Jx\rangle_\pi = x^\top (D_\pi J)x$ and $D_\pi J = \frac{1}{2}(D_\pi P - P^\top D_\pi)$ is antisymmetric. (ii) For $\|x\|_\pi = 1$, convexity gives $|(Px)_i|^2 \le \sum_j P_{ij}|x_j|^2$; weighting by π_i and using $\sum_i \pi_i P_{ij} = \pi_j$ gives $\|Px\|_\pi \le 1$, with equality at $x = \mathbf{1}$. The same holds for $P^\dagger$, and $\|S\|_\pi \le \frac{1}{2}(\|P\|_\pi + \|P^\dagger\|_\pi) = 1$. For the 4-cycle C, π is uniform and C is a permutation, hence normal; so is $J = \frac{1}{2}(C - C^{-1})$, whose norm is therefore its largest eigenvalue modulus, $\max|\frac{1}{2}(\zeta - \zeta^{-1})| = 1$ over the fourth roots of unity ζ. (iii) S is π-self-adjoint with norm 1 (T1, T3). If P is lazy then $P = \frac{1}{2}(I + P')$ with P' row-stochastic and π-stationary, so $S = \frac{1}{2}(I + S')$, where S' is the reversible part of P', with $\|S'\|_\pi \le 1$, and each eigenvalue is at least 0. □

Proof of Proposition 1.5, the tied limit. With $b = \sigma/(L+\sigma)$ the path length $K_L = |\mathcal{S}|$ is binomial(L,b) with mean $Lb = \sigma - \sigma^2/(L+\sigma)$. By T17, $d_{\mathrm{TV}}(K_L, K) \le Lb^2 + |Lb - \sigma| \le 2\sigma^2/L$ for $K \sim \mathrm{Poisson}(\sigma)$. Since every power of P has $\|P^k\|_\infty = 1$, $\|\mathbb{E}[P^{K_L}] - \mathbb{E}[P^K]\|_\infty = \|\sum_k(\Pr[K_L = k] - \Pr[K = k])P^k\|_\infty \le 2d_{\mathrm{TV}}(K_L,K) \le 4\sigma^2/L$, and $\mathbb{E}[P^K] = \sum_k e^{-\sigma}\sigma^k P^k/k! = e^{\sigma(P-I)}$. No diagonalizability is needed. □

The following statement underlies the token-side remark of Section 1.3; clause (i) is essentially the contraction argument of Oono & Suzuki (2020) and Cai & Wang (2020).

Proposition A.1 (Collapse is a sample-side property)**.** *Let P be reversible, irreducible and lazy, with spectral gap $\gamma = 1 - \mu_2 > 0$. Consider the stack $X^{(l+1)} = \varphi(PX^{(l)}W^{(l)})$ with $\|W^{(l)}\|_2 \le c_W$ and an entrywise 1-Lipschitz nonlinearity φ, and let $d_\pi(X) = \|(I - \mathbf{1}\pi^\top)X\|_\pi$ be the π-distance to the consensus set $\{\mathbf{1}c^\top\}$. Then*

(i) *$d_\pi(X^{(l+1)}) \le c_W(1-\gamma)\, d_\pi(X^{(l)})$, so the stack converges to the consensus set geometrically whenever $c_W(1-\gamma) < 1$;*

(ii) *with no assumption on P, $\operatorname{rank}(P^L XW) \le \operatorname{rank}(P^L X)$ for every W.*

Proof of Proposition A.1. Propagation contracts. $P\mathbf{1} = \mathbf{1}$ and $\pi^\top P = \pi^\top$ give $(I - \mathbf{1}\pi^\top)P = P(I - \mathbf{1}\pi^\top)$, so the part of X off the consensus direction is propagated by P restricted to $\mathbf{1}^{\perp_\pi}$, the span of $v_2, \dots, v_N$, where its norm is $\max_{k\ge 2}|\mu_k| = \mu_2 = 1 - \gamma$ by laziness (T1). Column by column, $d_\pi(PX) \le (1-\gamma)d_\pi(X)$. *Feature maps rescale.* $(I - \mathbf{1}\pi^\top)(XW) = ((I - \mathbf{1}\pi^\top)X)W$ and, row-wise, $\|YW\|_\pi^2 = \sum_i \pi_i \|W^\top y_i\|^2 \le \|W\|_2^2 \|Y\|_\pi^2$. *Nonlinearities do not expand.* $d_\pi(X)$ is the π-distance from X to the consensus set, attained at the projection $\mathbf{1}c^\top$ (Pythagoras in the

π-geometry). Since $\varphi(\mathbf{1}c^\top) = \mathbf{1}\varphi(c)^\top$ is itself a consensus matrix and φ is entrywise 1-Lipschitz, $d_\pi(\varphi(X)) \le \|\varphi(X) - \varphi(\mathbf{1}c^\top)\|_\pi \le \|X - \mathbf{1}c^\top\|_\pi = d_\pi(X)$. Chaining gives (i); (ii) is T10. □

Clause (ii) is nearly vacuous at finite depth, where only numerical rank collapses; clause (i) is the content: learned feature maps can slow collapse only by growing their norm.

A.3 Proofs for Chapter 2

The random-time representation (Section 2.1). Let (X_k) be a Markov chain with transition matrix P and let K be independent of it, with $\mathbb{P}(K = k) = p_k$ and possibly $\mathbb{P}(K = \infty) = p_\infty$. For every state i and set A,

$$\mathbb{P}_i(X_K \in A,\, K < \infty) = \sum_{k \ge 0} p_k \mathbb{P}_i(X_k \in A) = e_i^\top \Big(\sum_k p_k P^k\Big)\mathbf{1}_A.$$

The series converges absolutely because $\|P^k\|_\infty = 1$, so $f(P)$ is a nonnegative matrix with row sums $1 - p_\infty$. A stationary row satisfies $\pi^\top f(P) = (1 - p_\infty)\pi^\top$; for a proper law ($p_\infty = 0$), stationary laws are preserved and an absorbing state stays absorbing. For $Pv = \mu v$, $f(P)v = \sum_k p_k \mu^k v = f(\mu)v$; no normality is used. Independence of K from the walk is what makes the coefficients a single scalar sequence: stopping rules that depend on the visited states weight paths by their history and generally do not produce a scalar filter.

When P is reversible with respect to π and irreducible, $D_\pi^{1/2} P D_\pi^{-1/2}$ is symmetric, and in the π-weighted norm $\|f(P) - \mathbf{1}\pi^\top\|_\pi = \max_{j \ge 2} |f(\mu_j)|$ for a proper law (with mass p_∞ at infinity it is $\max(p_\infty, \max_{j \ge 2} |f(\mu_j)|)$, since the consensus direction has eigenvalue $-p_\infty$); if P is also lazy, all eigenvalues are nonnegative and $\max_{j \ge 2} |f(\mu_j)| = f(\mu_2)$ because f is nondecreasing on $[0, 1]$. The symmetric graph operator H of Section 2.7 is similar to the walk $\widetilde{D}^{-1}(W + I)$, so every scalar statement transfers to it.

Complements to Theorem 2.1. The bounds (2.2) and the equivalence are proved with the theorem. Four complements.

Sharpness. The law $q\delta_0 + (1 - q)\delta_{m+1}$ has $q_m = q$ and gain $q + (1 - q)\mu^{m+1}$, so the upper bound is attained. The lower bound is approached by $q\delta_m + (1 - q)\delta_n$ as $n \to \infty$ and attained by the defective law $q\delta_m + (1 - q)\delta_\infty$.

A certificate. If $f(\mu) \ge \gamma$ and m satisfies $\mu^{m+1} \le \gamma/2$, then $q_m \ge \gamma/(2 - \gamma)$.

Equivalent forms. For a family $\{f_\theta\}$ the following are equivalent: (i) bounded-window mass; (ii) a monomial floor $f_\theta(x) \ge c\, x^m$ for all $x \in [0, 1]$ and all θ; (iii) $\inf_\theta f_\theta(x) > 0$ for every $x \in (0, 1)$; (iv) $\inf_\theta f_\theta(x) > 0$ for one $x \in (0, 1)$. Indeed (i)⇒(ii) by the lower bound, (ii)⇒(iii)⇒(iv) trivially, and (iv)⇒(i) by the theorem. The floors in (iii) depend on x.

The mean. Jensen's inequality for the convex map $k \mapsto \mu^k$ gives $f(\mu) \ge \mu^{\mathbb{E}K}$, so a bounded mean suffices. The persistent-branch family $\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{2L-1}$ has mean L and gain $\frac{1}{2}\mu + \frac{1}{2}\mu^{2L-1} \to \mu/2$; the shrinking-branch family $\frac{1}{L}\delta_1 + (1 - \frac{1}{L})\delta_{L+1}$ has mean L and gain $\frac{1}{L}\mu + (1 - \frac{1}{L})\mu^{L+1} \to 0$. Neither has an identity atom.

Proof of Corollary 2.2. (a) Let λ be an eigenvalue of P with $|\lambda| < 1$ and let $\mathcal{G}_\lambda$ be its generalized eigenspace, on which $P = \lambda I + \mathcal{N}$ with $\mathcal{N}^r = 0$. Then $P^k = \sum_{j<r} \binom{k}{j} \lambda^{k-j} \mathcal{N}^j$ on $\mathcal{G}_\lambda$, and

$$f(P)\big|_{\mathcal{G}_\lambda} = \sum_{j<r} \mathbb{E}\Big[\tbinom{K}{j} \lambda^{K-j}\Big] \mathcal{N}^j,$$

with $\binom{k}{j} = 0$ for $k < j$ and the term at $K = \infty$ equal to zero. Each $g_j(k) = \binom{k}{j}|\lambda|^{k-j}$ is bounded and tends to zero as $k \to \infty$. If $K_n \to \infty$ in probability, then for every m, $\mathbb{E}g_j(K_n) \le \sup_{k>m} g_j(k) + \|g_j\|_\infty \mathbb{P}(K_n \le m)$, and letting $n \to \infty$ and then $m \to \infty$ gives $\mathbb{E}g_j(K_n) \to 0$. Hence $f_n(P)x \to 0$ on $\mathcal{G}_\lambda$, and on the sum of all such spaces. For irreducible aperiodic P, the eigenvalue 1 is simple with spectral projection $\mathbf{1}\pi^\top$ and every other eigenvalue has modulus below one, so $f_n(P) \to \mathbf{1}\pi^\top$ for proper laws.

(b) $|f(-\mu)| \le \sum_k p_k \mu^k = f(\mu)$ by the triangle inequality, and $f(\mu) - f(-\mu) = 2\sum_{k \text{ odd}} p_k \mu^k$, which vanishes if and only if $p_k = 0$ for every odd k. □

Proof of Proposition 2.3. If $f = \exp\{-\lambda(1-h)\}$, then $g_n = \exp\{-(\lambda/n)(1-h)\}$ is the generating function of a compound Poisson law on $\mathbb{N}_0$—a Poisson(λ/n) number of independent jumps with generating function h—and $g_n^n = f$. Conversely, suppose $f = g_n^n$ for every n with g_n generating functions of laws on $\mathbb{N}_0$. If $p_0 = 0$ then $g_n(0)^n = 0$, so each of the n summands is at least one and $K \ge n$ almost surely for every n, which is impossible for a law on $\mathbb{N}_0$. Hence $p_0 > 0$, the law is infinitely divisible on $\mathbb{N}_0$, and by Feller's theorem (Feller, 1968, Chapter XII.2) it is compound Poisson: $f(z) = \exp\{-\lambda(1-h(z))\}$ with $\lambda = -\log p_0$ and h the generating function of a law on $\{1, 2, \dots\}$. For $\mu \in [0,1]$ and $J \ge 1$, $0 \le h(\mu) = \mathbb{E}\mu^J \le \mu$, which gives $e^{-\lambda} \le f(\mu) \le e^{-\lambda(1-\mu)}$. For a family, $\sup_\theta \lambda_\theta = \Lambda < \infty$ gives $f_\theta(\mu) \ge e^{-\Lambda}$; if instead $\lambda_{\theta_n} \to \infty$, then $f_{\theta_n}(\mu) \le e^{-\lambda_{\theta_n}(1-\mu)} \to 0$. Finally $\mathbb{P}(K = 0) = e^{-\lambda}$.

The examples of Section 2.2 follow from their generating functions. Heat diffusion: $e^{t(z-1)}$, so $h(z) = z$. Fractional diffusion: $e^{-t(1-z)^\alpha}$ with $h(z) = 1 - (1-z)^\alpha$, whose coefficients $b_k = \alpha\Gamma(k-\alpha)/[\Gamma(1-\alpha)\Gamma(k+1)]$ are positive and sum to one (the Sibuya law, Section 2.1). Geometric and negative binomial: $[(1-q)/(1-qz)]^\nu = \exp\{-\lambda(1-h(z))\}$ with $\lambda = -\nu\log(1-q)$ and the logarithmic-series law $h(z) = \log(1-qz)/\log(1-q)$. A Bernoulli layer $1 - p + pz$ with $0 < p < 1$ has no square root with nonnegative coefficients (the second coefficient of $\sqrt{1-p+pz}$ is negative), so binomial laws and fixed $K = L \ge 1$ are not refinable. □

Proof of Corollary 2.4. $f(\mu_1)/f(\mu_2) = \sum_k p_k \mu_1^k / f(\mu_2) = \sum_k \widetilde{p}_k (\mu_1/\mu_2)^k$ with $\widetilde{p}_k = p_k \mu_2^k / f(\mu_2)$. For the uniform law on $\{0, \dots, L\}$ and $\mu_2 < 1$, $\widetilde{p}_0 = (1-\mu_2)/(1-\mu_2^{L+1}) \ge 1 - \mu_2$, and the ratio is at least $\widetilde{p}_0$. □

Effective path length. Tilting also identifies which path lengths carry a given mode. With $\widetilde{\mathbb{E}}_\mu$ the expectation under $p_k\mu^k/f(\mu)$,

$$\frac{d\log f(\mu)}{d\log \mu} = \widetilde{\mathbb{E}}_\mu K, \qquad \frac{d^2 \log f(\mu)}{d(\log\mu)^2} = \widetilde{\mathrm{Var}}_\mu K \ge 0.$$

Binomial(L,p) tilts to Binomial$(L, p\mu/(1-p+p\mu))$, Poisson(t) to Poisson$(t\mu)$, and a geometric law with ratio $1-r$ to one with ratio $(1-r)\mu$. The tilted mean is finite even when $\mathbb{E}K = \infty$. For Binomial$(100, \frac{1}{2})$ it is 23.1, 41.2 and 48.7 at $\mu = 0.3$, 0.7 and 0.95: the modal signal is carried by paths much shorter than the nominal mean, as Veit et al. (2016) observed for gradients.

Proof of Proposition 2.5 and of the mixing-time bound (2.3). Splitting at d, $f(\mu) \le \mathbb{P}(K < d) + \mu^d \mathbb{P}(K \ge d) = 1 - (1-\mu^d)\mathbb{P}(K \ge d)$, so $\mathbb{P}(K \ge d) \le (1-f(\mu))/(1-\mu^d) \le (1-\gamma)/(1-\mu^d)$. The law $(1-q)\delta_0 + q\delta_d$ with $q = (1-\gamma)/(1-\mu^d) \le 1$ has gain exactly γ and $\mathbb{P}(K \ge d) = q$. If node j is at graph distance d from i, then $(P^k)_{ij} = 0$ for $k < d$ and $(P^k)_{ij} \le 1$ otherwise, so $f(P)_{ij} \le \mathbb{P}(K \ge d)$. For the Sibuya jump law $\mathbb{P}(S > k) = \prod_{j=1}^k (1-\alpha/j) = \Gamma(k+1-\alpha)/[\Gamma(1-\alpha)\Gamma(k+1)] \sim k^{-\alpha}/\Gamma(1-\alpha)$; the jumps are subexponential, so the compound Poisson sum satisfies $\mathbb{P}(K \ge d) \sim t\, d^{-\alpha}/\Gamma(1-\alpha)$

(Embrechts et al., 1997, Section 1.3). The numerical check at $t = 1$, $\alpha = \frac{1}{2}$ and $d = 2000$ agrees to four digits.

For (2.3) and a proper path law, row i of $f(P) - \mathbf{1}\pi^\top$ is $\sum_k p_k(e_i^\top P^k - \pi^\top)$, whose ℓ_1 norm is at most

$$\sum_k p_k \, 2d_{\mathrm{TV}}(P^k(i,\cdot),\pi) \le 2\mathbb{E}\, d(K).$$

Since d is nonincreasing and at most one, $\mathbb{E}\, d(K) \le \mathbb{P}(K < t(\varepsilon)) + \varepsilon$. □

Proof of Theorem 2.8(iii) and (iv). Restated for completeness from Bruce (2026k, App. A.2) (the complements to Theorem 2.1 above). The engine is a cascade: *if $f_\theta(\mu_0) \to 0$ along a subfamily for some $\mu_0 \in (0,1)$, then $f_\theta(\mu) \to 0$ along it for every $\mu \in (0,1)$.* Indeed, the lower truncation bound gives $\Pr[K_\theta \le m] \le \mu_0^{-m} f_\theta(\mu_0) \to 0$ for each fixed m, and the upper bound then gives $\limsup_\theta f_\theta(\mu) \le \mu^m$ for every m, hence 0. (ii)⇒(iii): for $\mu \in [0,1]$, $f_\theta(\mu) \ge \sum_{k\le m} p_\theta(k)\mu^k \ge \mu^m \Pr[K_\theta \le m] \ge \rho\mu^m$. (iii)⇒(i): evaluate the floor at μ_C. (i)⇒(iv): (i) gives (ii), hence (iii), hence $f_\theta(\mu_0) \ge \rho\mu_0^m > 0$ at every interior μ_0. (iv)⇒(i): if $\kappa_{\mathrm{col},\theta} \to \infty$ along a subfamily then $f_\theta(\mu_C) \to 0$ and the cascade sends $f_\theta(\mu_0) \to 0$ at every interior μ_0. □

Proof of Theorem 2.8, products clause. Restated from Bruce (2026k, Thm. 6 and App. B.1) (Theorem 2.11 and the complements to it below), whose mixing budget $\sum_\ell \Pr[K_\ell \ge 1]$ is the defect $\sum_\ell(1-\hat a_\ell)$ here. For hops of one operator, the composed filter acts on each eigenvector by $\prod_\ell f_\ell(\mu)$ (T1), so $\ln \kappa_{\mathrm{col},L} = -\sum_\ell \ln f_\ell(\mu_C)$. Each hop satisfies $f_\ell(\mu_C) \ge \hat a_\ell$ (nonnegative coefficients) and, being convex on $[0,1]$ with $f_\ell(0) = \hat a_\ell$ and $f_\ell(1) = 1$, $f_\ell(\mu_C) \le 1 - \beta(1-\hat a_\ell)$. *Necessity:* $-\ln f_\ell(\mu_C) \ge \beta(1-\hat a_\ell)$, so bounded $\kappa_{\mathrm{col},L}$ forces $\sum_\ell(1-\hat a_\ell) < \infty$. *Sufficiency:* summability forces $\hat a_\ell \ge \frac{1}{2}$ for all but finitely many ℓ, where $-\ln f_\ell(\mu_C) \le -\ln \hat a_\ell \le 2(1-\hat a_\ell)$; the tail is summable and the finitely many remaining factors are positive. The clause assumes one operator: with a different spectral gap at every hop the two directions separate, leaving a band of cases the anchor defect alone does not decide. □

Proof of Proposition 2.9. As in Section 2.3, $f(z) = \sum_k c_k z^k$ with $f(1) = 1$, and $n_- = \sum_k (-c_k)_+$ is its negative coefficient mass, so that $\sum_k |c_k| = 1 + 2n_-$. (i) The positive coefficients have mass $1 + n_-$, and $\mu_0^k \in [\mu_0^L, 1]$ for $0 \le k \le L$, so $\mu_0^L - n_-(1-\mu_0^L) \le f(\mu_0) \le 1 + n_-(1-\mu_0^L)$; this is the lower bound. For $\varrho < \mu_0^L$ the filter $(1+n_-)z^L - n_-$ with $n_- = (\mu_0^L - \varrho)/(1-\mu_0^L)$ attains it, for $\varrho > 1$ the filter $(1+n_-) - n_- z^L$ with $n_- = (\varrho - 1)/(1-\mu_0^L)$, and for $\varrho \in [\mu_0^L, 1]$ a convex combination of 1 and z^L has no negative mass. (ii) Write $V = \sum_k |c_k|$, so that $|f(w)| \le V$ on the circle $|w| = R$. Let p interpolate f at the $m \ge 2$ Chebyshev extrema x_j of $[0, a_0]$. The interpolation weights at $\mu_C > a_0$ alternate in sign like the values of the Chebyshev polynomial T_{m-1} at its extrema, so their absolute sum is $T_{m-1}(z) \le \Lambda^{m-1}$, and $|p(\mu_C)| \le \epsilon_N \Lambda^{m-1}$. Hermite's formula for the remainder, with $\omega(w) = \prod_j (w - x_j)$,

$$f(\mu_C) - p(\mu_C) \;=\; \frac{1}{2\pi i} \oint_{|w|=R} \frac{f(w)\,\omega(\mu_C)}{(w-\mu_C)\,\omega(w)}\, dw,$$

together with $|\mu_C - x_j|/|w - x_j| \le \mu_C/(R - a_0) = r$, gives $|f(\mu_C) - p(\mu_C)| \le RVr^m/(R-\mu_C)$. Hence $1 - \epsilon_C \le \epsilon_N \Lambda^{m-1} + RVr^m/(R-\mu_C)$ for every $m \ge 2$. With $m = 1 + \lfloor \ln((1-\epsilon_C)/(2\epsilon_N))/\ln \Lambda \rfloor$, which the hypothesis makes at least 2, the first term is at most $(1-\epsilon_C)/2$ and $r^{-m} \ge ((1-\epsilon_C)/(2\epsilon_N))^\gamma$; rearranging gives the bound. (iii) The lower bound is (ii). For the upper, let $k = \lceil L(\mu_N + \mu_C)/2 \rceil$ and $f_L(x) = \sum_{j \ge k} \binom{L}{j} x^j (1-x)^{L-j} = \Pr[\mathrm{Bin}(L,x) \ge k]$. It has $f_L(1) = 1$ and values in $[0,1]$, and by Hoeffding's inequality $f_L \le e^{-L(\mu_C - \mu_N)^2/2}$ on $[0, \mu_N]$ and $1 - f_L \le e^{-L(\mu_C-\mu_N)^2/2}$ on $[\mu_C, 1]$.

Expanding in monomials, $\sum_k |c_k| \le \sum_{j\ge k} \binom{L}{j} 2^{L-j} \le 3^L$, and the stated degree makes both band errors at most ϵ_N with $3^L \le 3\epsilon_N^{-K}$. □

Proof of the Poisson tail bound that follows Proposition 2.10. By T14 with $t = \ln(m/\sigma) \ge 0$ and $\mathbb{E}e^{tK} = e^{\sigma(e^t-1)}$, $\Pr[K \ge m] \le e^{-\sigma}(e\sigma/m)^m$ for $m \ge \sigma$; below σ the bound fails (at $\sigma = 4$, $m = 1$ it gives 0.20 against the true 0.98). A geometric lottery has $\Pr[K \ge m] = \alpha^m$. □

Proof of Corollary 2.12, with complements to Theorem 2.11. The proof of Theorem 2.11 uses only $(1-\mu)\mathbb{P}(K_\ell \ge 1) \le 1 - f_\ell(\mu) \le \mathbb{P}(K_\ell \ge 1)$ and $1 - x \le -\log x \le (1-x)/x$. For Bernoulli layers with probability p_ℓ, $f_\ell(\mu) = 1 - p_\ell(1-\mu) \ge \mu$; for the energy layers of Theorem 2.13, $B_\ell \le 2$ gives $f_\ell(\mu) \ge \mu^2$. With $\rho_L \ge \rho > 0$ the two sides of (2.6) are within the factor $1/(\rho(1-\mu))$, which proves the four regimes of Corollary 2.12. When $p_\ell \to 0$ the lower bound is asymptotically exact: $-\log\prod_\ell(1-(1-\mu)p_\ell) = (1-\mu)S_L + O(\sum_\ell p_\ell^2)$.

The critical case. For $p_\ell = c/\ell$ with $0 < c \le 1$ and $a = c(1-\mu)$, repeated use of $\Gamma(x+1) = x\Gamma(x)$ gives $\prod_{\ell=1}^{L}(1-a/\ell) = \Gamma(L+1-a)/[\Gamma(1-a)\Gamma(L+1)]$, and Stirling's formula gives $\sim L^{-a}/\Gamma(1-a)$. For $c = \frac{1}{2}$ and $\mu = 0.85$ the gain is 0.674 at $L = 100$ and 0.338 at $L = 10^6$: a slowly declining curve without a floor.

Fixed sequences. For one infinite sequence of layer laws, the gain has a positive limit if and only if $\sum_\ell \mathbb{P}(K_\ell \ge 1) < \infty$. Necessity is the lower bound. For sufficiency, the tail of the sum is small, so $a_\ell = \mathbb{P}(K_\ell = 0) > \frac{1}{2}$ for $\ell > \ell_0$; then $\prod_{\ell>\ell_0} f_\ell(\mu) \ge \prod_{\ell>\ell_0} a_\ell \ge \exp\{-2\sum_{\ell>\ell_0}(1-a_\ell)\} > 0$, and the finitely many earlier factors are positive. Equivalently, by the Borel–Cantelli lemmas, only finitely many layers mix almost surely (Durrett, 2019).

Depth families. If every layer of an L-layer network has mixing probability p_L, then $S_L = Lp_L$ and a floor holds if and only if $\sup_L Lp_L < \infty$; if $Lp_L \to c$ the gain tends to $e^{-c(1-\mu)}$, and if $p_L \sim cL^{-\eta}$ with $0 < \eta < 1$ its negative logarithm is asymptotic to $c(1-\mu)L^{1-\eta}$. Without the condition on ρ_L a bounded budget does not suffice: one layer with $K_\ell = L$ has budget one and gain μ^L.

Nonlinear residual GCNs. Let H be the augmented normalized adjacency of a graph, $\mathcal{U}$ the span of its top eigenvectors (one per connected component, each with nonnegative entries), $\mathcal{M} = \mathcal{U} \otimes \mathbb{R}^C$, and $d_{\mathcal{M}}(X)$ the Frobenius distance from $X \in \mathbb{R}^{N\times C}$ to $\mathcal{M}$; let $\lambda < 1$ be the largest modulus of the remaining eigenvalues. Oono & Suzuki (2020) show that $d_{\mathcal{M}}(HX) \le \lambda\, d_{\mathcal{M}}(X)$, $d_{\mathcal{M}}(XW) \le \|W\|_2\, d_{\mathcal{M}}(X)$ and $d_{\mathcal{M}}(\mathrm{ReLU}(X)) \le d_{\mathcal{M}}(X)$. Because $d_{\mathcal{M}}$ is a seminorm, the update $X_\ell = (1-p_\ell)X_{\ell-1} + p_\ell\,\mathrm{ReLU}(HX_{\ell-1}W_\ell)$ with $\|W_\ell\|_2 \le 1$ satisfies $d_{\mathcal{M}}(X_\ell) \le (1-p_\ell)d_{\mathcal{M}}(X_{\ell-1}) + p_\ell\lambda\, d_{\mathcal{M}}(X_{\ell-1}) = f_\ell(\lambda)\, d_{\mathcal{M}}(X_{\ell-1})$, where $f_\ell(z) = 1 - p_\ell + p_\ell z$ is the layer's generating function. Iterating and using $1 - x \le e^{-x}$ gives $d_{\mathcal{M}}(X_L) \le e^{-(1-\lambda)S_L} d_{\mathcal{M}}(X_0)$ with $S_L = \sum_\ell p_\ell$. The bound is one-sided, so it transfers escape but not retention. □

Proof of Theorem 2.13. With $X_0 = vw_0^\top$ and $Pv = \mu v$, induction gives $X_\ell = vw_\ell^\top$ with $w_\ell^\top = w_{\ell-1}^\top(I + \alpha_\ell\mu V_\ell)$. Write $\alpha = \alpha_\ell$, $m = m_\ell$, $V = V_\ell = mI + Z$ and $w = w_{\ell-1}$. Then

$$\|w_\ell\|^2 = w^\top(I+\alpha\mu V)(I+\alpha\mu V^\top)w = \|w\|^2 + \alpha\mu\, w^\top(V+V^\top)w + \alpha^2\mu^2\, w^\top VV^\top w.$$

Conditionally on $\mathcal{F}_{\ell-1}$, $\mathbb{E}[V+V^\top] = 2mI$ and $\mathbb{E}[VV^\top] = m^2I + m\,\mathbb{E}[Z+Z^\top] + \mathbb{E}[ZZ^\top] = (m^2+\sigma_\ell^2)I$, so $\mathbb{E}[\|w_\ell\|^2 \mid \mathcal{F}_{\ell-1}] = g_\ell(\mu)\|w\|^2$. The tower property gives $G_L(\mu) = \prod_\ell g_\ell(\mu)$, and dividing by the case $\mu = 1$ gives R_L. For $\mu \in [0,1]$ and $\alpha_\ell m_\ell \ge 0$, $g_\ell(z)/g_\ell(1)$ has nonnegative coefficients summing to one, so it is the generating function of $B_\ell \in \{0,1,2\}$ as stated, and $R_L(\mu) = \mathbb{E}[\mu^{\sum_\ell B_\ell}]$ by independence of the factors.

For the depth family, $S_L = L\,(2\alpha_L m_L + \alpha_L^2(m_L^2+\sigma_L^2))/g(1)$ and $\rho_L \ge \mu^2$, so by Theorem 2.11 a floor holds if and only if S_L is bounded. If S_L is bounded then $\alpha_L m_L \to 0$ and $\alpha_L^2\sigma_L^2 \to 0$, so

$g(1) \to 1$, $L\alpha_L^2 m_L^2 = (L\alpha_L m_L)\,\alpha_L m_L \to 0$, and boundedness of S_L is equivalent to (2.7). With $m = 0$ and $\alpha_L = c/\sqrt{L}$, $R_L \to e^{-c^2\sigma^2(1-\mu^2)}$; with $m > 0$ fixed and the same α_L, expanding $\log g$ gives $-\log R_L = 2cm(1-\mu)\sqrt{L} + O(1)$.

Anisotropy. If $\sigma_-^2 I \preceq \mathbb{E}[Z_\ell Z_\ell^\top \mid \mathcal{F}_{\ell-1}] \preceq \sigma_+^2 I$, the same computation gives

$$\prod_\ell \frac{(1+\alpha_\ell m_\ell \mu)^2 + \alpha_\ell^2 \sigma_-^2 \mu^2}{(1+\alpha_\ell m_\ell)^2 + \alpha_\ell^2 \sigma_+^2} \;\le\; R_L(\mu) \;\le\; \prod_\ell \frac{(1+\alpha_\ell m_\ell \mu)^2 + \alpha_\ell^2 \sigma_+^2 \mu^2}{(1+\alpha_\ell m_\ell)^2 + \alpha_\ell^2 \sigma_-^2},$$

so the sufficiency direction of (2.7) holds under any bounded anisotropy, and necessity holds whenever $\sigma_+^2 \mu^2 < \sigma_-^2$.

Layer-index schedules. With $m = 0$ and $\alpha_\ell = c/\sqrt{\ell}$, $R_L = \prod_{\ell \le L} (\ell + c^2\sigma^2\mu^2)/(\ell + c^2\sigma^2)$, which is the Gamma ratio in Section 2.4; for $c = \sigma = 1$, $\mu = 0.9$ and $L = 10^4$ it equals 0.186, while $\alpha_L = 1/\sqrt{L}$ gives 0.827. □

Proof of Proposition 2.14. Exact formula. On the mode, one visit maps w to $U_j w$ with $U_j = I + \varepsilon\mu V_j^\top = (1 + a\mu s_j \lambda) I - a\mu s_j \beta J$, where $a = \varepsilon\sigma$. Since $J + J^\top = 0$ and $JJ^\top = I$, $U_j U_j^\top = [(1 + a\mu s_j\lambda)^2 + a^2\mu^2\beta^2] I = (1 + 2a\mu s_j \lambda + a^2\mu^2) I$, so U_j is a multiple of an orthogonal matrix and $\|U_j^N w\|^2 = (1 + 2a\mu s_j\lambda + a^2\mu^2)^N \|w\|^2$ for every w. Averaging over $s_j = \pm 1$ gives $G_N(a\mu)$ per block, independence across blocks gives the M-th power, and dividing by $\mu = 1$ gives $R_{M,N}$. The formula agrees with direct enumeration of all sign patterns on explicit two-dimensional feature vectors to 10^{-12} (Bruce, 2026k, App. H).

Lower bound. Write $G_N(u) = (1+u^2)^N E_N(x(u))$ with $x(u) = 2\lambda u/(1+u^2) \in [0,1]$ and $E_N(x) = \frac{1}{2}[(1+x)^N + (1-x)^N] = \sum_k \binom{N}{2k} x^{2k}$. Then

$$\log R_{M,N}(\mu) = MN \log \frac{1 + a^2\mu^2}{1 + a^2} - M\big[\log E_N(x(a)) - \log E_N(x(a\mu))\big].$$

The first term is at least $-MNa^2(1-\mu^2)$. For the second, the coefficient inequality $(2k+2)\binom{N}{2k+2} \le N(N-1)\binom{N}{2k}$, equivalent to $(N-2k)(N-2k-1) \le N(N-1)(2k+1)$, gives $E_N'(x) \le N(N-1)\,x\,E_N(x)$, hence $\log E_N(x_1) - \log E_N(x_0) \le \frac{1}{2}N(N-1)\max(0, x_1^2 - x_0^2)$. Now $x(u)^2 = 4\lambda^2 g(u)$ with $g(u) = u^2/(1+u^2)^2$ and $g'(u) = 2u(1-u^2)/(1+u^2)^3 \le 2u$, so $g(a) - g(a\mu) \le a^2(1-\mu^2)$. Together,

$$\log R_{M,N}(\mu) \;\ge\; -MNa^2(1-\mu^2) - 2MN(N-1)\lambda^2 a^2(1-\mu^2) = -MN\varepsilon^2\sigma^2(1-\mu^2)\big[1 + 2(N-1)\lambda^2\big].$$

At $\lambda = 1$ this is the bound $\exp\{-MN(2N-1)a^2(1-\mu^2)\}$ for aligned maps; at $\lambda = 0$ it is the exact rotational law's elementary bound. Substituting $\varepsilon = c/(N\sqrt{M})$ gives exponent at least $-c^2\sigma^2(1-\mu^2)(2N-1)/N \ge -2c^2\sigma^2(1-\mu^2)$, and $\varepsilon = c/\sqrt{MN}$ gives $-c^2\sigma^2(1-\mu^2)(1 + 2(N-1)\lambda^2)$.

Loss of the floor. Take $\varepsilon = c/\sqrt{MN}$ and write $\kappa = c\sigma$, so that $a = \kappa/\sqrt{MN}$; let $MN \ge 4\kappa^2$, so that $a \le \frac{1}{2}$. For $0 \le u \le \frac{1}{2}$ and $\lambda \in [0,1]$, $|\log(1 \pm 2\lambda u + u^2) \mp 2\lambda u| \le 5u^2$, so $G_N(u)$ lies between $e^{-5Nu^2}\cosh(2N\lambda u)$ and $e^{5Nu^2}\cosh(2N\lambda u)$: $\log G_N(u) = \log\cosh(2N\lambda u) + O(Nu^2)$ uniformly in λ. Since $MNa^2 = \kappa^2$,

$$\Big| -\log R_{M,N}(\mu) - M\big[\log\cosh z - \log\cosh \mu z\big] \Big| \;\le\; 10\kappa^2, \qquad z = 2\kappa\sqrt{N\lambda^2/M},$$

for every M, N and λ. For $\mu \in [0,1)$ the function $h(z) = \log\cosh z - \log\cosh \mu z$ has $h'(z) = \tanh z - \mu\tanh\mu z \ge (1-\mu)\tanh z$, so $h(z) \ge (1-\mu)\log\cosh z \ge c_\mu \min(z^2, z)$ with $c_\mu = (1-\mu)\log\cosh 1$, because $\log\cosh z/z^2$ decreases and $\log\cosh z/z$ increases on $(0,\infty)$. Hence

$$-\log R_{M,N}(\mu) \;\ge\; c_\mu \min\big(4\kappa^2 N\lambda^2,\; 2\kappa\sqrt{M \cdot N\lambda^2}\big) - 10\kappa^2,$$

which diverges as $N\lambda^2 \to \infty$ for every $M \geq 1$, uniformly in M. For fixed $\lambda \in (0,1]$ and fixed M, $z \to \infty$ and $\log\cosh z = z - \log 2 + o(1)$, so the two-sided bound gives $-\log R_{M,N}(\mu) = M(1-\mu)z + O(1) = 2c\sigma\lambda(1-\mu)\sqrt{MN} + O(1)$, the stated asymptotic. The checks of Bruce (2026k, App. H) confirm the lower bound for M up to 10^{30} and the constant to 5×10^{-4} at $N = 10^6$. □

Proof of Proposition 2.15. (a) Token 1 is absorbing. Each $i \geq 2$ moves to token 1 with probability at least $P_{i1} > 0$ at every visit, so it is visited finitely often, and the stationarity equations on the columns $j \geq 2$ read $\pi_T^\top Q = \pi_T^\top$; since every row of Q sums to less than one, $\|\pi_T^\top Q^k\|_1 \leq (\max_i \sum_j Q_{ij})^k \|\pi_T\|_1 \to 0$, so $\pi_T = 0$ and $\pi = e_1$. *(b)* Stationarity of π_ε rearranges to $\pi_\varepsilon^\top = \frac{\varepsilon}{N}\mathbf{1}^\top(I - (1-\varepsilon)P)^{-1}$ (T4 in the ℓ_∞ norm), the PageRank identity (Page et al., 1999). Expanding the series and reading a column $j \geq 2$, where only the transient block contributes (T7), $\pi_\varepsilon(j) = \frac{\varepsilon}{N}\sum_{i\geq 2}\sum_k (1-\varepsilon)^k (Q^k)_{ij}$. With $F = (I-Q)^{-1}$ the fundamental matrix, whose entries count expected visits before absorption (Kemeny & Snell, 1976), the discount changes each term $\sum_k (Q^k)_{ij}$ by at most $\sum_k k\varepsilon(1-\delta)^k = \varepsilon(1-\delta)/\delta^2$, so $\pi_\varepsilon(j) = \frac{\varepsilon}{N}\big(\sum_{i\geq 2} F_{ij}\big)(1+O(\varepsilon))$, with $\sum_{i\geq 2} F_{ij} \in [1, N/\delta]$. Hence $\pi_\varepsilon(1) = 1 - O(\varepsilon)$, the smallest entry is $\Theta(\varepsilon)$, and $\sqrt{\pi_{\max}/\pi_{\min}} = \Theta(\varepsilon^{-1/2})$ for fixed P. □

Proof of Theorem 2.16(i)–(iii). Parts (i) and (ii) restate the single-kernel case of Bruce (2026k, Thm. 11) (Section 8.3). By induction, $P^k = \left(\begin{smallmatrix} 1 & 0 \\ \mathbf{1} - Q^k\mathbf{1} & Q^k \end{smallmatrix}\right)$, and $(Q^k\mathbf{1})_i = \Pr[\tau_i > k]$ is the probability that the walk from i is still unabsorbed after k steps. Averaging over the lottery, row $i \geq 2$ of $f(P) - \mathbf{1}e_1^\top$ is $(-(\mathbb{E}[Q^K]\mathbf{1})_i,\ \mathbb{E}[Q^K]_{i,\cdot})$: a nonpositive first entry followed by nonnegative entries. Their absolute values therefore add with no cancellation, and the row's ℓ_1 size is $2(\mathbb{E}[Q^K]\mathbf{1})_i = 2\Pr[\tau_i > K]$, which is (i). For (ii), each row of Q sums to at most $1-\delta$, so $(Q^k\mathbf{1})_i \leq (1-\delta)^k$; for self-or-sink this is an equality. For (iii), $\tau_i \geq 1$ gives $\Pr[K < \tau_i] \geq \Pr[K = 0] = a$; the class-extremal value is (ii) together with its equality case, and the equivalence is Lemma 2.6(c) at $c = 1-\delta$. □

Proof of Theorem 2.16(iv). Q is triangular, so its eigenvalues are its diagonal entries (T7), $\rho(Q) = \max_{i\geq 2} P_{ii}$, and $\|Q^L\|_\infty^{1/L} \to \rho(Q)$ (T19). Monotonicity: Q is nonnegative with $Q\mathbf{1} \leq (1-\delta)\mathbf{1}$, so $Q^{k+1}\mathbf{1} = Q^k(Q\mathbf{1}) \leq Q^k\mathbf{1}$ entrywise, and $\|Q^k\|_\infty = \max_i (Q^k\mathbf{1})_i$ is nonincreasing and at most $(1-\delta)^k$. □

Proof of Theorem 2.18. (i) Row i of $f(P)$ is the mixture $\sum_k p_k (P^k)_{i\cdot}$, the law $\mathcal{L}_i$ of the walk from i stopped at the independent time K, and $\|f(P) - \mathbf{1}\nu^\top\|_\infty = 2\max_i d_{\mathrm{TV}}(\mathcal{L}_i, \nu)$. Stationarity gives $\nu = \nu P^k$, so by T9 applied k times, $d_{\mathrm{TV}}((P^k)_{i\cdot}, \nu) = d_{\mathrm{TV}}(e_i^\top P^k, \nu P^k) \leq (1-\lambda)^k d_{\mathrm{TV}}(e_i, \nu) = (1-\lambda)^k(1-\nu_i)$; convexity over the mixture (T8) gives $d_{\mathrm{TV}}(\mathcal{L}_i, \nu) \leq (1-\nu_i)\mathbb{E}[(1-\lambda)^K]$. *(ii)* On the two-community chain with $h \leq \frac{1}{2}$, $f(P) - \mathbf{1}\pi^\top = f(1-2h)\, v_2 v_2^\top D_\pi$ with $v_2 = (1,-1)$, whose rows each have ℓ_1 size $f(1-2h)$, while the right side of (i) is $2 \cdot \frac{1}{2} \cdot \mathbb{E}[(1-2h)^K]$; the two rows overlap in $2h = \gamma$, which is therefore the leak. For strictly causal P with $\delta > 0$, $\nu = e_1$ (Proposition 2.15) and $\lambda = \delta$, so (i) reads $2\mathbb{E}[(1-\delta)^K]$, which self-or-sink attains (Theorem 2.16(ii)). *(iii)* The upper bound is (i) with $1 - \nu_{\min} \leq 1$ and $\lambda \geq \lambda_0$; self-or-sink at $\delta = \lambda_0$ has leak λ_0 and attains it; the equivalence is Lemma 2.6(c) at $c = 1 - \lambda_0$. □

Proof of Proposition 2.19. (i) is the first step of the proof of Theorem 2.18; at the causal end $d_{\mathrm{TV}}(\mu, e_1) = 1 - \mu(\{1\})$ and $\mathcal{L}_i(\{1\}) = \Pr[\tau_i \leq K]$. (ii) By T9, $1 - \lambda(AB) \leq (1-\lambda(A))(1-\lambda(B))$, so $1 - \lambda(P^k) \leq (1-\lambda(P^m))^{\lfloor k/m \rfloor}$; run the proof of Theorem 2.18(i) with this rate. (iii) If $Pv = \zeta v$ with $\zeta \neq 1$, then $\nu^\top v = \nu^\top P v = \zeta \nu^\top v$ forces $\nu^\top v = 0$, so $(f(P) - \mathbf{1}\nu^\top)v = f(\zeta)v$ and the norm is at least $|f(\zeta)|$. For the anchor floor, $\mathcal{L}_i(\{i\}) \geq \Pr[K = 0] = a$, so $d_{\mathrm{TV}}(\mathcal{L}_i, \nu) \geq a - \nu_i$ (T8); take i with $\nu_i = \nu_{\min}$. (iv) Write row i of P^k as the density $\chi^{(k)} = D_\pi^{-1}(P^\top)^k e_i$ against π; then $2d_{\mathrm{TV}}((P^k)_{i\cdot}, \pi) = \sum_j \pi_j |\chi_j^{(k)} - 1| \leq \|\chi^{(k)} - \mathbf{1}\|_\pi$ (T2), and $\chi^{(k)} - \mathbf{1} = (P^\dagger)^k(e_i/\pi_i - \mathbf{1})$, a vector

in $\mathbf{1}^{\perp_\pi}$ of initial size $\sqrt{1/\pi_i - 1}$ acted on by $(P^\dagger)^k$, whose restricted norm equals that of P^k (T3); average over the lottery. □

The constant of (iv) grows like the conditioning of Proposition 2.15(b), $\Theta(\varepsilon^{-1/2})$, as the regularization of a causal operator is removed, so the π-metric law, continued toward the causal end, pays as its constant the same order of degeneracy that the causal law avoids by working in ℓ_∞.

Proof of Proposition 2.20(i) and (iii). (i) is the scalar-share case of Bruce (2026k, Thm. 11) (Section 8.3), restated for completeness. Each factor is strictly causal, with transient block $(1-b_\ell)I+b_\ell Q_\ell$ whose rows sum to at most $1-b_\ell\delta_\ell$ and whose diagonal is at least $1-b_\ell$. The transient block of F is the product of these blocks, and the distance is $2\max_{i\geq 2}$ of its row sums (as in Theorem 2.16); submultiplicativity gives the upper bound, and the diagonal of a product of lower-triangular matrices is the product of the diagonals, which gives the lower. (iii) Write row i of P_ℓ as $\theta_\ell e_i + (1-\theta_\ell)q_i$ with q_i a probability vector (possible because $(P_\ell)_{ii} \geq \theta_\ell$). For the pair (i,j) attaining $\operatorname{diam}(X_\ell)$, $x'_i - x'_j = (1-b_\ell+b_\ell\theta_\ell)(x_i-x_j) + b_\ell(1-\theta_\ell)(q_iX_\ell - q_jX_\ell)$, and the last difference has norm at most $\operatorname{diam}(X_\ell)$ because both terms are convex combinations of rows of X_ℓ. □

Proof of Proposition 2.20(ii). A factor $(1-b)I+bP$ has uniform leak at least $b\lambda(P)$: two of its rows overlap at least as much as b times the overlap of the corresponding rows of P. By T9, $1-\lambda(F) \leq \prod_\ell(1-b_\ell\lambda(P_\ell))$, and since every factor fixes ν, $d_{\mathrm{TV}}(e_i^\top F,\nu) = d_{\mathrm{TV}}(e_i^\top F,\nu F) \leq (1-\lambda(F))(1-\nu_i)$. □

Proof of Proposition 2.21, and the other placements. Model. Let the token-mixing branch be $X \leftarrow X + P\mathcal{N}(X)V_\ell$ and the token-wise branch $X \leftarrow X + \mathcal{N}(X)W_\ell$, with V_ℓ, W_ℓ centered and isotropic as in Theorem 2.13, and let $\mathcal{N}$ rescale the stream so that each branch writes a fixed expected energy, v and v', into the consensus channel. Per-token LayerNorm and RMSNorm do this up to fluctuations of the per-token norms; replacing the random normalizing constant by its expectation is the mean-field approximation, exact when per-token norms are equal and accurate to leading order in width when they concentrate (Section 2.6 measures the residual finite-width effect). The token-mixing branch is then the energy-law layer with $\alpha_\ell^2\sigma^2 = v/E_{\ell-1}$, where $E_{\ell-1}$ is the consensus energy before it. The token-wise branch has $P = I$, so it multiplies every mode's expected energy, the consensus included, by the same factor and cancels in R_L; it only adds v' to E. Hence $E_\ell = E_0 + \ell(v+v')$ and

$$R_L(\mu) = \prod_{\ell=1}^{L} \frac{1+\mu^2 v/E_{\ell-1}}{1+v/E_{\ell-1}} = \prod_{\ell=1}^{L} \frac{e+\ell-1+s\mu^2}{e+\ell-1+s}.$$

Using $\prod_{\ell=1}^{L}(x+\ell-1) = \Gamma(x+L)/\Gamma(x)$ gives the Gamma form, and $\Gamma(x+\alpha)/\Gamma(x+\beta) \sim x^{\alpha-\beta}$ gives $L^{-s(1-\mu^2)}$. Layer ℓ takes a step with probability $q_\ell/(1+q_\ell)$, $q_\ell = s/(e+\ell-1)$, and $\sum_{aL<\ell\leq bL} q_\ell/(1+q_\ell) \to s\ln(b/a)$.

Coherent Pre-LN. With $V_\ell = \sigma I$ (pure drift) and normalization by the consensus amplitude, each branch adds a fixed amplitude, the relative weight of layer ℓ is $\asymp s'/\ell$ with s' the attention share of amplitude, and the amplitude ratio is $\prod_\ell(e'+\ell-1+s'\mu)/(e'+\ell-1+s') \asymp L^{-s'(1-\mu)}$; its square, the energy ratio, decays as $L^{-2s'(1-\mu)}$, the slope tested in Figure 2.2c.

LayerNorm Scaling. Multiplying the normalized input of layer ℓ by $1/\sqrt{\ell}$ scales both increments by $1/\ell$, so $E_\ell = E_0 + (v+v')H_\ell$ with H_ℓ the harmonic number, $q_\ell \sim s/(\ell\log\ell)$, the budget grows as $s\log\log L$, and $-\log R_L = s(1-\mu^2)\log\log L + O(1)$.

Depth-scaled branches. Multiplying both branches by $c/\sqrt{L}$ adds $c^2(v+v')/L$ per layer, so $E_\ell \leq E_0 + c^2(v+v')$, $q_\ell \leq c^2 v/(LE_0)$, and the budget is at most c^2v/E_0: a floor at every depth.

Post-LN. Normalizing after the addition resets the stream energy to a fixed E^*, so $q_\ell = v/E^*$ is constant and $R_L = [(1+\mu^2 q)/(1+q)]^L$; with $v = E^*$ the per-layer factor is $(1+\mu^2)/2$, the reference value in Section 2.6. □

The four-state example (Section 2.8). Let $B(\mu) = \frac{1}{2}\left(\begin{smallmatrix} 1+\mu & 1-\mu \\ 1-\mu & 1+\mu \end{smallmatrix}\right)$ and $P = B(\mu_s) \otimes B(\mu_n)$. The vectors $v_s = \frac{1}{2}(1,1,-1,-1)$ and $v_n = \frac{1}{2}(1,-1,1,-1)$ are orthonormal eigenvectors with eigenvalues μ_s and μ_n. For input $x = v_s + av_n$, target $y = v_s$ and identity readout,

$$\frac{\|f(P)x - y\|^2}{\|x\|^2} = \frac{(f(\mu_s)-1)^2 + a^2 f(\mu_n)^2}{1+a^2}.$$

With $\mu_s = \frac{19}{20}$, $\mu_n = \frac{3}{10}$, $a = \frac{3}{2}$, $L = 12$, the lazy residual $(\frac{1}{2} + \frac{z}{2})^{12}$ and the restart law with $r = \frac{1}{5}$, exact rational arithmetic gives signal gains 0.7380 and 0.8395, nuisance gains 0.0057 and 0.2632, and errors 0.0211 and 0.0559; plain powers give 0.0650. The example refutes a monotone relation between retention and fixed-readout error; it does not estimate how often the reversal occurs.

The two-mode frontier, several caps and the continuous envelope (Proposition 2.22). For path laws on $\mathbb{N}_0 \cup \{\infty\}$ the achievable pairs $(f(\mu_-), f(\mu_+))$ form the closed convex hull of the points (μ_-^k, μ_+^k), $k \in \mathbb{N}_0$, together with $(0,0)$. These points lie on $y = x^q$ with $q = \log \mu_+ / \log \mu_- \in (0,1)$, which is concave and increasing, so the upper boundary of the hull is the piecewise-linear interpolation U through consecutive points; the proof of Proposition 2.22 uses exactly this. Replacing U by the curve itself gives the continuous envelope $f(\mu_+) \le f(\mu_-)^q$, with equality if and only if K is almost surely constant, by strict concavity of x^q. At $(\mu_-, \mu_+, \eta) = (0.5, 0.9, 0.01)$ the frontier value is 0.493177 and the continuous envelope 0.496585.

With caps $f(\mu_j) \le \eta_j$ for $j = 1, \dots, J$, maximizing $f(\mu_+)$ over laws on $\{0, \dots, K_{\max}\}$ is a linear program with $J+1$ constraints besides nonnegativity, and a basic optimal solution has at most $J+1$ nonzero coordinates. Negative eigenvalues enter as two-sided caps $-\eta \le f(\mu) \le \eta$. In our checks with two and three caps, the optimum was always an adjacent pair of depths chosen by the binding cap (Bruce, 2026k, App. H).

A.4 Proofs for Chapter 3

Proof of Theorem 3.1. Perron vectors. $\widetilde{P}u = D_\pi^{1/2} P \mathbf{1} = D_\pi^{1/2}\mathbf{1} = u$ and $u^\top \widetilde{P} = \pi^\top P D_\pi^{-1/2} = \pi^\top D_\pi^{-1/2} = u^\top$; the same holds for $\widetilde{P}^\top$ and hence for $\widetilde{S}$, while $\widetilde{J}u = \frac{1}{2}(u-u) = 0$. With $\widetilde{E} = \widetilde{S} - uu^\top$ we get $\widetilde{E}u = u - u(u^\top u) = 0$, since $u^\top u = \sum_i \pi_i = 1$.

Orthogonality. Symmetric and skew matrices are Frobenius-orthogonal, which settles the pairs $(uu^\top, \widetilde{J})$ and $(\widetilde{E}, \widetilde{J})$. For the remaining pair, $\langle uu^\top, \widetilde{E}\rangle_F = u^\top \widetilde{E} u = 0$. Since $\|uu^\top\|_F^2 = (u^\top u)^2 = 1$, we get $\|\widetilde{P}\|_F^2 = 1 + \|\widetilde{E}\|_F^2 + \|\widetilde{J}\|_F^2$.

Trace identity. $\|\widetilde{S}\|_F^2 - \|\widetilde{J}\|_F^2 = \frac{1}{4}(\|\widetilde{P} + \widetilde{P}^\top\|_F^2 - \|\widetilde{P} - \widetilde{P}^\top\|_F^2) = \langle \widetilde{P}, \widetilde{P}^\top\rangle_F = \sum_{ij} \widetilde{P}_{ij}\widetilde{P}_{ji} = \sum_{ij} \sqrt{\frac{\pi_i}{\pi_j}} P_{ij} \sqrt{\frac{\pi_j}{\pi_i}} P_{ji} = \operatorname{tr}(P^2)$. Adding $\|\widetilde{S}\|_F^2 + \|\widetilde{J}\|_F^2 = \|\widetilde{P}\|_F^2$ gives $2\|\widetilde{J}\|_F^2 = \|\widetilde{P}\|_F^2 - \operatorname{tr}(P^2)$, which is (3.4). Since $\operatorname{tr}(P^2) = \sum_{ij} P_{ij}P_{ji} \ge 0$, we have $\kappa_{\mathrm{circ}} \le 1$, with equality iff every product $P_{ij}P_{ji}$ vanishes. Also $\kappa_{\mathrm{circ}} = 0$ iff $\widetilde{J} = 0$ iff $J = 0$.

Permutation kernels. If P is the matrix of a permutation σ, then π is constant on each cycle of σ for any stationary law, so $P^\dagger = P^\top$ and $\widetilde{P} = P$. Then $\|\widetilde{P}\|_F^2 = N$, and $\operatorname{tr}(P^2) = \#\{i : \sigma^2(i) = i\}$ counts the tokens on cycles of length one or two.

Nearest self-adjoint operator. An operator R is self-adjoint in $L^2(\pi)$ iff $\widetilde{R}$ is symmetric. For symmetric Y, $\|\widetilde{P} - Y\|_F^2 = \|\widetilde{S} - Y\|_F^2 + \|\widetilde{J}\|_F^2$, which is minimized uniquely at $Y = \widetilde{S}$.

Circulation information. Write $D(R\|Q) = \sum_i \mu_i D_{\mathrm{KL}}(R_{i\cdot}\|Q_{i\cdot})$, with μ the stationary law of R, for the stationary divergence rate (Section 10.2), so that $\mathcal{I}(P) = D(P\|S)$. Since $P_{ij}/S_{ij} = F_{ij}/A_{ij}$, $D(P\|S) = \sum_{ij} F_{ij}\log(F_{ij}/A_{ij})$ and $D(P^\dagger\|S) = \sum_{ij} F_{ji}\log(F_{ji}/A_{ij})$. Relabeling $i \leftrightarrow j$ and using the symmetry of A shows these are equal. Because $S_{i\cdot} = \frac{1}{2}(P_{i\cdot} + P^\dagger_{i\cdot})$, averaging the two gives

$\sum_i \pi_i \,\mathrm{JS}(P_{i\cdot}, P_{i\cdot}^{\dagger}) \le \log 2$. Convexity of D_{KL} in its second argument gives $D_{\mathrm{KL}}(p\|\frac{p+q}{2}) \le \frac{1}{2}D_{\mathrm{KL}}(p\|q)$, hence $D(P\|S) \le \frac{1}{2}D(P\|P^{\dagger}) = \frac{1}{2}\sum_{ij} F_{ij}\log(F_{ij}/F_{ji}) = \sigma/2$. Finally, $D(P\|S) = -H_\pi(P) - \sum F \log S$, and $\sum (F-A)\log S = \frac{1}{2}\sum_{ij} C_{ij}(\log A_{ij} - \log \pi_i) = 0$. The first term vanishes because C is skew and A symmetric, and the second because $C\mathbf{1} = 0$. So $\sum F \log S = \sum A \log S = -H_\pi(S)$, which gives $D(P\|S) = H_\pi(S) - H_\pi(P)$. □

Proof of Theorem 3.2. (a) For antisymmetric a let $\psi_j = \frac{1}{N}\sum_i a_{ij}$. Antisymmetry gives $\frac{1}{N}\sum_j a_{ij} = -\psi_i$ and $\mathbf{1}^\top a \mathbf{1} = 0$, so $(HaH)_{ij} = a_{ij} - \psi_j + \psi_i$, that is, $a = (\mathbf{1}\psi^\top - \psi\mathbf{1}^\top) + HaH$. The matrix $r = HaH$ has zero row and column sums. Any potential difference $\mathbf{1}\phi^\top - \phi\mathbf{1}^\top$ is orthogonal to it, because $\langle \mathbf{1}\phi^\top, r\rangle = \sum_j \phi_j \sum_i r_{ij} = 0$ and likewise for the other term. Potential differences span an $(N-1)$-dimensional subspace of the $\frac{N(N-1)}{2}$-dimensional space of antisymmetric matrices, and doubly centered antisymmetric matrices span its orthogonal complement, so the split is unique. Then $s = \frac{1}{2}(s + s^\top) + \frac{1}{2}a$, and the term $-\frac{1}{2}\psi\mathbf{1}^\top$ is constant along rows and cancels in softmax.

(b) With $P_{ij} = e^{s_{ij}}/z_i$, detailed balance $\pi_i P_{ij} = \pi_j P_{ji}$ reads $s_{ij} - s_{ji} = \phi_j - \phi_i$ with $\phi = \log\pi - \log z$ ($\Rightarrow$). Conversely, if $a_{ij} = \psi_j - \psi_i$ then $\pi_i \propto z_i e^{\psi_i}$ satisfies detailed balance ($\Leftarrow$). Hence $J = 0$ iff a is a potential difference, iff $r = 0$ by (a). For the cycle affinity, $\log(P_{ij}/P_{ji}) = a_{ij} - \log z_i + \log z_j$; around a cycle the normalizer terms telescope, and so do the potential terms of a.

(c) If $s = \lambda QK^\top + x\mathbf{1}^\top + \mathbf{1}y^\top$ then $a = \lambda(QK^\top - KQ^\top) + \mathbf{1}(y-x)^\top - (y-x)\mathbf{1}^\top$, and $H\mathbf{1} = 0$ gives $r = HaH = \lambda(HQ)(HK)^\top - \lambda(HK)(HQ)^\top$. No property of Q, K is used, so rotary features and untied projections are covered. For $s = XWX^\top$, write $W = W_{\mathrm{sym}} + W_{\mathrm{sk}}$ to get $a = 2XW_{\mathrm{sk}}X^\top$ and $r = 2X_c W_{\mathrm{sk}} X_c^\top$. This vanishes iff $x^\top W_{\mathrm{sk}} y = 0$ for all x, y in the span of the centered rows; for collinear centered rows it vanishes because $x^\top W_{\mathrm{sk}} x = 0$.

(d) $\sigma = \sum_{ij} F_{ij}\log(F_{ij}/F_{ji}) = \frac{1}{2}\sum_{ij} C_{ij}\log(F_{ij}/F_{ji})$ because the log-ratio is antisymmetric. With $\log(F_{ij}/F_{ji}) = (\phi_i - \phi_j) + a_{ij}$ and $C\mathbf{1} = C^\top\mathbf{1} = 0$, the potential terms drop out, leaving $\sigma = \frac{1}{2}\langle C, a\rangle = \frac{1}{2}\langle C, s\rangle - \frac{1}{2}\langle C, s^\top\rangle = \langle C, s\rangle$, using $\langle C, s^\top\rangle = \langle C^\top, s\rangle = -\langle C, s\rangle$. The potential part of a also drops out of $\langle C, a\rangle$, so $\sigma = \frac{1}{2}\langle C, r\rangle$. For $s = \lambda QK^\top$, $\langle C, s\rangle = \lambda \operatorname{tr}(C^\top QK^\top) = \lambda\langle Q, CK\rangle$. □

Proof of Proposition 3.3. (a) $f^\top D_\pi J f = \frac{1}{2} f^\top C f = 0$ since C is skew. The Dirichlet identity follows from A being symmetric with row sums π.

(b) For unit $x \in \mathbb{C}^N$, $x^*\widetilde{S}x \in \mathbb{R}$ and $x^*\widetilde{J}x \in i\mathbb{R}$ ($\widetilde{J}$ is real skew), so $\mathrm{Re}\, x^*(\widetilde{S} + \alpha\widetilde{J})x = x^*\widetilde{S}x$. The subspace $u^\perp$ is invariant under $\widetilde{P}_\alpha = \widetilde{S} + \alpha\widetilde{J}$ because $\widetilde{P}_\alpha^\top u = u$. It carries the $N-1$ nontrivial eigenvalues, and on it the Rayleigh quotient of $\widetilde{S}$ lies in $[\lambda_{\min}(\widetilde{S}), \lambda_2(\widetilde{S})]$. For the semigroup, take $f \perp_\pi \mathbf{1}$ and $f_t = e^{t(P_\alpha - I)}f$. Then $\frac{d}{dt}\|f_t\|_\pi^2 = 2\langle f_t, (S-I)f_t\rangle_\pi \le -2\,\mathrm{gap}(S)\|f_t\|_\pi^2$ by (a) and the Poincaré inequality, and the process stays mean-zero.

(c) $P_\alpha X_0 = P_\alpha X - \mathbf{1}\pi^\top X$ and $JX_0 = JX$. Expanding, $\|P_{\pm a}X_0\|_\pi^2 = \|SX_0\|_\pi^2 \pm 2a\langle SX_0,\, JX_0\rangle_\pi + a^2\|JX\|_\pi^2$; averaging removes the cross term. In stationary coordinates the cross term is

$$\operatorname{tr}(\widetilde{X}_0^\top \widetilde{S}\widetilde{J}\widetilde{X}_0) = \tfrac{1}{2}\operatorname{tr}(\widetilde{X}_0^\top[\widetilde{S}, \widetilde{J}]\widetilde{X}_0),$$

since only the symmetric part of $\widetilde{S}\widetilde{J}$ contributes. Finally, $\widetilde{P}\widetilde{P}^\top - \widetilde{P}^\top\widetilde{P} = -2[\widetilde{S}, \widetilde{J}]$, and $[\widetilde{S}, \widetilde{J}]$ annihilates u.

(d) A circulant P is doubly stochastic, so π is uniform and $P^\dagger = P^\top$, whose kernel is $k(-d)$. Hence S and J have kernels $\frac{1}{2}(k(d) \pm k(-d))$. The discrete Fourier transform diagonalizes all three, with real transfer function for S and imaginary transfer function for J. For $k = \delta_{+1}$, $Sf = \frac{1}{2}(f_{i+1} + f_{i-1})$ and $Jf = \frac{1}{2}(f_{i+1} - f_{i-1})$. □

In the remaining proofs, $P_s = S + sJ$ is the sweep (1.3) at $t = s$, with $|s| \le 1$, so $\|P_s\|_\pi = 1$ (Lemma 1.2) and $\|J\|_\pi = \Phi$.

Proof of Theorem 3.4(ii). $F(s) = \sum_k p_k \langle Y, P_s^k Y\rangle_\pi$ is twice differentiable because $\mathbb{E}[K^2] < \infty$. Being even, it has $F'(0) = 0$. The second derivative of P_s^k is a sum of $k(k-1)$ words, one for each ordered pair of distinct positions at which a factor J replaces a factor P_s; each word has π-norm at most Φ^2. Hence $|F''(s)| \le \Phi^2 \mathbb{E}[K(K-1)]\|Y\|_\pi^2$ for $|s| \le 1$, and T12 gives the bound. □

Proof of Theorem 3.4(iii).

$$G'(0) = \sum_k p_k \sum_{a+b=k-1} \langle X, S^a J S^b Y\rangle_\pi.$$

In the eigenbasis of S, $\langle v_k, S^a J S^b v_l\rangle_\pi = \mu_k^a \mu_l^b J_{kl}$ (T1), and $\sum_{a+b=n-1} \mu_k^a \mu_l^b$ is the divided difference of z^n at (μ_k, μ_l); summing over the lottery gives $f[\mu_k, \mu_l] J_{kl}$. The divided difference is symmetric in its arguments and $J_{lk} = \langle v_l, Jv_k\rangle_\pi = \langle J^\dagger v_l, v_k\rangle_\pi = -J_{kl}$, so the array is antisymmetric, $\mathcal{D}_f^\dagger = -\mathcal{D}_f$, and $\langle X, \mathcal{D}_f Y\rangle_\pi = -\langle Y, \mathcal{D}_f X\rangle_\pi$. For a simple eigenvalue, T13 with $A = S$ and $B = J$ gives $\mu_j'(0) = J_{jj} = 0$ (by Lemma 1.2(i)) and the stated eigenvector derivative. □

Proof of Theorem 3.4(iv). $\frac{d}{ds}\|P_s Y\|_\pi^2 = 2\langle P_s Y, JY\rangle_\pi$, which at $s = 0$ is $\langle SY, JY\rangle_\pi + \langle JY, SY\rangle_\pi = \langle Y, (SJ - JS)Y\rangle_\pi$, using $S^\dagger = S$ and $J^\dagger = -J$. The commutator $[S, J]$ is π-self-adjoint, so its quadratic form vanishes for every real Y exactly when $[S, J] = 0$, and $PP^\dagger - P^\dagger P = -2[S, J]$. □

Proof of Proposition 3.6. $W(t)^\dagger = P_t^{(1)\dagger} \cdots P_t^{(L)\dagger}$, which is the reversed word at $-t$; quadratic readouts do not see the adjoint, which gives (i). For (ii), differentiate the product at $t = 0$; conservation (Lemma 1.2(i)) gives $\langle u, Jw\rangle_\pi = \frac{1}{2}\langle u - w, J(u + w)\rangle_\pi$, and Cauchy–Schwarz with $\|u_\ell\|_\pi, \|w_\ell\|_\pi \le \|Y\|_\pi$ gives the bound. □

Proof of Proposition 3.7(i). Let $V = \mathbf{1}^{\perp_\pi}$, which every P_s and J leave invariant, and let B_s be the restriction of $f(P_s)$ to V, so that $d_f(P_s) = \|B_s\|_\pi$ (the matrix $f(P_s) - \mathbf{1}\pi^\top$ vanishes on $\mathbf{1}$ and agrees with $f(P_s)$ on V). Let $\mathcal{D}$ be the restriction of $\mathcal{D}_f$ to V; it is π-antisymmetric with $\|\mathcal{D}\|_\pi \le \Phi\, \mathbb{E}K$. The word count in the proof of Theorem 3.4(ii), applied in operator norm (T12), gives $\|B_s - B_0 - s\mathcal{D}\|_\pi \le \frac{1}{2} s^2 \Phi^2 \mathbb{E}[K(K-1)]$, so the Gram operator $\mathcal{G}(s) = B_s^\dagger B_s$ satisfies

$$\mathcal{G}(s) = B_0^2 + s\,(B_0\mathcal{D} - \mathcal{D}B_0) + R(s), \qquad \|R(s)\|_\pi \le Cs^2 \quad (|s| \le 1),$$

with C depending only on Φ, $\mathbb{E}K$ and $\mathbb{E}[K(K-1)]$. Let $c = d_f(S) = \max_{k \ge 2} |f(\mu_k)| > 0$, let V_{top} be the span of the v_k $(k \ge 2)$ with $|f(\mu_k)| = c$, and let $c' < c$ be the largest $|f(\mu_k)|$ on its complement in V (or 0 if the complement is trivial). By hypothesis B_0 equals c on all of V_{top} or $-c$ on all of it; in either case, for $x_{\mathrm{top}} \in V_{\mathrm{top}}$ the two terms of $\langle x_{\mathrm{top}}, (B_0\mathcal{D} - \mathcal{D}B_0)x_{\mathrm{top}}\rangle_\pi$ are equal, and it vanishes. Write a unit $x \in V$ as $x_{\mathrm{top}} + x_\perp$. Every surviving part of the first-order term involves $x_\perp$, and $\|B_0\mathcal{D} - \mathcal{D}B_0\|_\pi \le 2c\,\Phi\,\mathbb{E}K$, so that term is at most $C'|s|\|x_\perp\|_\pi$ with $C' = 6c\,\Phi\,\mathbb{E}K$, and $\langle x, \mathcal{G}(s)x\rangle_\pi \le c^2 - (c^2 - c'^2)\|x_\perp\|_\pi^2 + C'|s|\|x_\perp\|_\pi + Cs^2 \le c^2 + (C + C'^2/(4(c^2 - c'^2)))s^2$. Conversely, a unit $x \in V_{\mathrm{top}}$ gives $\langle x, \mathcal{G}(s)x\rangle_\pi \ge c^2 - Cs^2$. Hence $d_f(P_s)^2 = c^2 + O(s^2)$ and, since $c > 0$, $d_f(P_s) = c + O(s^2)$. If P is lazy, every $\mu_k \in [0, 1]$ (Lemma 1.2(iii)) and $f \ge 0$ there, so the hypothesis holds. Without it, if $f(\mu_k) = c$ and $f(\mu_l) = -c$ at two top modes, the unit vector $x = (v_k + v_l)/\sqrt{2}$ has first-order coefficient $2c\,\langle v_k, \mathcal{D}_f v_l\rangle_\pi$ in $\langle x, \mathcal{G}(s)x\rangle_\pi$, which is generically nonzero; the witness of Section 3.5 realizes $d_f = 0.2 + \Phi$ exactly (a machine-precision check). □

Proof of Proposition 3.7(ii). Write the lazy P as $\frac{1}{2}(I + P')$ with P' row-stochastic and π-stationary. Then $S - P^\dagger P = \frac{1}{4}(I - P'^\dagger P')$, which is positive semidefinite in the π-geometry because $\|P'\|_\pi \le 1$. For $x \in V$, $\|Px\|_\pi^2 = \langle x, P^\dagger P x\rangle_\pi \le \langle x, Sx\rangle_\pi \le \mu_2 \|x\|_\pi^2$ (T1). So the restriction B of P has $\|B\|_\pi \le \sqrt{\mu_2}$, and $d_f(P) = \|f(B)\|_\pi \le \sum_k p_k \|B\|_\pi^k \le \mathbb{E}[\mu_2^{K/2}]$. For the lazy cycle $\frac{1}{2}(I + Z)$ on N

tokens, π is uniform, the operator is normal, and each eigenvalue $\zeta = \frac{1}{2}(1+e^{i\phi})$ has $|\zeta|^2 = \frac{1}{2}(1+\cos\phi)$, the matching eigenvalue of S; so $\|B^L\|_\pi = \mu_2^{L/2}$. □

Proof of Proposition 3.7(iii). If $PP^\dagger = P^\dagger P$ then $D_\pi^{1/2} P D_\pi^{-1/2}$ is a normal matrix, S and J commute and share an orthonormal eigenbasis, and each eigenvalue is $\zeta = \xi + i\eta$ with ξ an eigenvalue of S and $i\eta$ one of J, so $|\eta| \le \Phi$. The restriction B is normal, so $\|f(B)\|_\pi = \max|f(\zeta)|$ over its eigenvalues, which are those of P other than the simple eigenvalue 1 (T19). For $\xi \in (0,1)$, $|f(\xi+i\eta)|^2 = f(\xi+i\eta)f(\xi-i\eta)$ (real coefficients) is even in η, and its η^2 coefficient is $f'(\xi)^2 - f(\xi)f''(\xi) = -f(\xi)^2(\ln f)''(\xi)$. With $\xi = e^u$, $\frac{d}{du}\ln f = \mathbb{E}_\xi K$ and $\frac{d^2}{du^2}\ln f = \mathrm{Var}_\xi K$ for the tilted lottery, so $\xi^2(\ln f)''(\xi) = \mathrm{Var}_\xi K - \mathbb{E}_\xi K$, which gives the formula. For the Poisson lottery, $|e^{\sigma(\xi+i\eta-1)}| = e^{\sigma(\xi-1)}$. A deterministic lottery tilts to itself, a geometric one to a geometric one, and a Poisson one to a Poisson one; a lottery on $\{0, 10\}$ with probabilities $(0.01, 0.99)$ has $\mathbb{E}K - \mathrm{Var}\,K = +8.91$ but tilted value -7.16 at $\xi = 0.5$ (a machine-precision check). □

Proof of Proposition 3.7(iv). If K given T is Poisson(T), then $f(z) = \mathbb{E}_T[e^{T(z-1)}]$ and $f(B) = \mathbb{E}_T[e^{T(B-I)}]$. On V, $\mathrm{Re}\langle x, (B-I)x\rangle_\pi = \langle x, (S-I)x\rangle_\pi \le (\mu_2 - 1)\|x\|_\pi^2$ because $\langle x, Jx\rangle_\pi$ is imaginary (T11), so T15 gives $\|e^{T(B-I)}\|_\pi \le e^{-T\bar\gamma}$ and $d_f(P) \le \mathbb{E}_T[e^{-T\bar\gamma}] = f(\mu_2)$. For S, f is positive and increasing on $[-1, 1]$, so $d_f(S) = f(\mu_2)$. The restart is Poisson with an exponentially distributed mean, and the residual stream is Poisson with a fixed mean. For depth, the lazy cycle gives $d_f(P) = \mu_2^{L/2} > \mu_2^L = d_f(S)$, while random lazy operators with $d_f(P) < d_f(S)$ exist (ratio 0.67 at $L = 4$, in a synthetic check). □

Three statements to which Section 3.5 points are stated here, each before its proof.

Proposition A.2 (The flow is the half-height of the numerical range)**.** *Let P be row-stochastic with stationary $\pi > 0$ and write $W_\pi(P) = \{\langle x, Px\rangle_\pi : x \in \mathbb{C}^N,\ \|x\|_\pi = 1\}$. Then* $\mathrm{spec}(P) \subseteq W_\pi(P) \subseteq [\mu_N, 1] + i\,[-\Phi, \Phi]$*, where μ_N is the smallest eigenvalue of S; the real parts of $W_\pi(P)$ fill exactly $[\mu_N, 1]$, independent of J, and the imaginary parts fill exactly $[-\Phi, \Phi]$.*

Proof of Proposition A.2. For complex x, $\langle x, Jx\rangle_\pi$ is purely imaginary (T11), so $\mathrm{Re}\langle x, Px\rangle_\pi = \langle x, Sx\rangle_\pi$, which sweeps exactly the spectral interval of S (T1), and $\mathrm{Im}\langle x, Px\rangle_\pi = \mathrm{Im}\langle x, Jx\rangle_\pi$, which sweeps exactly $[-\|J\|_\pi, \|J\|_\pi]$ (T11). Every eigenvalue lies in $W_\pi(P)$: take x a unit eigenvector. □

Proposition A.3 (Spectral floor and a shear pair)**.** *For every P with stationary $\pi > 0$, $d_f(P) \ge \max\{|f(\zeta)| : \zeta \in \mathrm{spec}(P),\ \zeta \ne 1\}$. There are operators with the same π and the same spectrum whose collapse distances differ.*

Proof of Proposition A.3. If $Pv = \zeta v$ with $\zeta \ne 1$, then $\pi^\top v = \pi^\top P v = \zeta \pi^\top v$ forces $\pi^\top v = 0$, so $(f(P) - \mathbf{1}\pi^\top)v = f(\zeta)v$ and $d_f(P) \ge |f(\zeta)|$. For the pair, let u_1, u_2 be a π-orthonormal basis of $V = \mathbf{1}^{\perp_\pi}$ on three tokens and $P(B) = \mathbf{1}\pi^\top + [u_1\ u_2]\,B\,[u_1\ u_2]^\top D_\pi$, which acts as 1 on $\mathbf{1}$ and as B on V, so $d_f = \|f(B)\|_2$ in these coordinates. With $B_1 = qI$ and $B_2 = \left(\begin{smallmatrix} q & r \\ 0 & q \end{smallmatrix}\right)$, both operators have spectrum $\{1, q, q\}$, while $f(B_2) = \left(\begin{smallmatrix} f(q) & rf'(q) \\ 0 & f(q) \end{smallmatrix}\right)$ has norm larger than $f(q)$ whenever $rf'(q) \ne 0$. At $\pi = (0.2, 0.3, 0.5)$, $q = 0.15$, $r = 0.05$ both operators are entrywise positive, and the restart $f(z) = 0.3/(1 - 0.7z)$ gives 0.3352 and 0.3418 (a machine-precision check). □

Proposition A.4 (Projector drift)**.** *For a simple eigenvalue μ_j of S with gap $\gamma_j = \min_{l \ne j}|\mu_j - \mu_l|$ and $\Phi < \gamma_j/2$, the spectral projector of P for the eigenvalues inside the circle $\Gamma = \{|z - \mu_j| = \gamma_j/2\}$ satisfies $\|\mathrm{Proj}_\Gamma(P) - \mathrm{Proj}_\Gamma(S)\|_\pi \le 2\Phi/(\gamma_j - 2\Phi)$.*

Proof of Proposition A.4. Write $\mathrm{Proj}_\Gamma(M) = \frac{1}{2\pi i}\oint_\Gamma (z-M)^{-1}dz$; the resolvent identity gives

$$\mathrm{Proj}_\Gamma(P) - \mathrm{Proj}_\Gamma(S) = \frac{1}{2\pi i}\oint_\Gamma (z-P)^{-1}J(z-S)^{-1}dz.$$

On Γ, $\|(z-S)^{-1}\|_\pi = 2/\gamma_j$ and, by T4, $\|(z-P)^{-1}\|_\pi \le 1/(\gamma_j/2-\Phi)$; the circle has length $\pi\gamma_j$. □

The first is Bendixson–Hirsch localization read in the walk's own metric (Horn & Johnson, 1991): Φ is how far the operator's action leaves the real axis, and on the lopsided triangle (Worked example 1.7) the operator never stops being a contraction while its eigenvalues collide and leave the axis. The bound of the third blows up as $\Phi \to \gamma_j/2$, where eigenvalues can collide: in a machine-precision check on the lopsided triangle, the collision at $s = 0.05$ occurs at $\Phi = \sqrt{3}\cdot 0.05 = 0.0866$, exactly half the gap $0.3366 - 0.1634$.

Proof of Lemma 3.8. With L the strictly lower part of M, $2M_- = L - L^\top$ and $2M_+^{\mathrm{off}} = L + L^\top$; since L and $L^\top$ have disjoint supports, both have squared norm $\|L\|_F^2 + \|L^\top\|_F^2$. The diagonal of M_+ is the diagonal of P, unchanged by conjugation. □

A.5 Proofs for Chapter 4

Proof of Theorem 4.1. (i) By Cauchy–Schwarz the maximum over unit $z \perp_\pi y$ is attained at $z^\star$, parallel to the escape r, and equals $\|r\|_\pi$ (if $r = 0$, both sides vanish). (ii) Expanding in the eigenbasis, $\|g(P)y\|_\pi^2 = \|y\|_\pi^2 \mathbb{E}_w[g^2]$ and $\langle y,\, g(P)y\rangle_\pi = \|y\|_\pi^2 \mathbb{E}_w[g]$, so $\|r\|_\pi^2 = \|g(P)y\|_\pi^2 - \langle y,\, g(P)y\rangle_\pi^2/\|y\|_\pi^2 = \|y\|_\pi^2\, \mathrm{sd}_w(g)^2$. (iii) $Ex = \varepsilon\langle z^\star,\, x\rangle_\pi u$, so $\|E\|_\pi = \varepsilon$, $Ey = 0$ because $z^\star \perp_\pi y$, and $\|Eg(P)y\|_\pi = \varepsilon|\langle z^\star,\, g(P)y\rangle_\pi| = \varepsilon\|r\|_\pi$. For any E with $Ey = 0$, $g(P)y = \mathrm{Proj}_y g(P)y + r$ and the first term is a multiple of y, so $Eg(P)y = Er$ and $\|Eg(P)y\|_\pi \le \|E\|_\pi\|r\|_\pi$. (iv) $M_n y = y$, so $Y_{\mathrm{post}} = g(P)y$, while $\langle n,\, M_n x\rangle_\pi = 0$ for every x; the maximum over n is (i). □

Proof of Theorem 4.3. Expand $G = \prod_a(\Lambda_a + H_a)$ into 2^A ordered words, indexed by the set $\mathcal{S}$ of filters that contribute their hollow part. The word with $\mathcal{S} = \emptyset$ contributes $\prod_a\langle v_j,\, G_a v_j\rangle_\pi$ at mode j, since the Λ_a are diagonal in the readout basis. A word with a single hollow factor H_a has the form $\Lambda' H_a \Lambda''$ with Λ', Λ'' diagonal; diagonal matrices and their adjoints map v_j to multiples of v_j, so its (j,j) element is a multiple of $\langle v_j,\, H_a v_j\rangle_\pi = 0$. A word with $|\mathcal{S}| \ge 2$ has (j,j) element at most its norm, which is at most $\prod_{a\in\mathcal{S}} h_a$ because $\|\Lambda_a\|_\pi = \max_k |\langle v_k,\, G_a v_k\rangle_\pi| \le \|G_a\|_\pi \le 1$. Summing over $|\mathcal{S}| \ge 2$ gives $\prod_a(1+h_a) - 1 - \sum_a h_a$, and $\prod_a(1+h_a) \le e^x$ with $x = \sum_a h_a$ together with $e^x - 1 - x \le \frac{1}{2}x^2 e^x$ gives the second bound. For $A = 2$ the only word with $|\mathcal{S}| \ge 2$ is H_2H_1. For the consensus mode, $G_a\mathbf{1} = \mathbf{1}$ gives $\Lambda_a\mathbf{1} = \mathbf{1}$ and $H_a\mathbf{1} = 0$; in any word containing a hollow factor, the rightmost one acts on a multiple of $\mathbf{1}$ and returns zero, so $\Delta_1 = 0$. For the witness, two polynomials in one reversible operator commute, and relative to a readout basis that is not that operator's eigenbasis their hollow parts are nonzero. Concretely, on four tokens with uniform π, a restart of a symmetric doubly stochastic operator and the operator itself commute to 5.6×10^{-17}, and read in the eigenbasis of a second such operator with simple spectrum they deviate by 0.024 at a readout mode (a machine-precision check). □

A.6 Proofs for Chapter 5

Proof of Proposition 5.1. The fixed point. A row-stochastic P satisfies $\|PX\|_{\infty,2} \le \|X\|_{\infty,2}$ in the maximum-row norm, so the map $X \mapsto (1-\alpha)X_0 + \alpha PX$ contracts by $\alpha < 1$. Its unique fixed point is the Neumann series $(1-\alpha)\sum_{k\ge 0}\alpha^k P^k X_0 = RX_0$, and every row of R sums to one.

Part (a). The cosines of output row i with the unit inputs are the entries of $(RG)_{i\cdot}$, divided by the row's norm. Normalizing multiplies them all by the same positive number, and a strict winner implies a nonzero row. For orthonormal rows $G = I$.

Part (b). On a cycle of length m, the resolvent puts weight $(1-\alpha)\alpha^r/(1-\alpha^m)$ on the token at cyclic distance $r = 0, \dots, m-1$, the geometric series over returns. The own weight, $r = 0$, is the largest for every $\alpha < 1$. A token d steps before the cycle has own weight $1-\alpha$, weight $(1-\alpha)\alpha^r$ on the path token $r < d$ steps ahead, and weight $(1-\alpha)\alpha^{d+r}/(1-\alpha^m)$ on the cycle token at distance r. The own weight beats every path token's, since $\alpha^r < 1$. Among the cycle tokens the entry point's weight, $(1-\alpha)\alpha^d/(1-\alpha^m)$, is the largest, and the own weight beats it exactly when $1-\alpha^m > \alpha^d$.

Part (c). Starting from the token's own input, write d_t for its own coefficient minus its partner's. Then $d_0 = 1$ and $d_{t+1} = 1-\alpha-\alpha d_t$, so $d_t = [1-\alpha+2\alpha(-\alpha)^t]/(1+\alpha)$, which is positive at every even t and, at odd t, exactly when $1-\alpha > 2\alpha^{t+1}$. □

Near-orthogonal inputs. With coherence $\eta = \max_{i\neq j}|G_{ij}| < 1$, the difference of the two correlations in (a) is $(R_{ii}-R_{ij})(1-G_{ij}) + \sum_{k\neq i,j} R_{ik}(G_{ki}-G_{kj})$. Since the row sums to one, it is at least $(R_{ii}-R_{ij})(1-\eta) - 2\eta(1-R_{ii}-R_{ij})$. So $R_{ii} > R_{ij}$ together with $(R_{ii}-R_{ij})(1-\eta) > 2\eta(1-R_{ii}-R_{ij})$ for every $j \neq i$ is sufficient. Identical inputs cannot be decoded uniquely whatever R is.

Two-cycles. For a symmetric, strict, self-excluding nearest-neighbour score, a directed cycle $i_1 \to i_2 \to \cdots \to i_m \to i_1$ with $m \geq 3$ would need $s(i_1,i_2) < s(i_2,i_3) < \cdots < s(i_m,i_1) < s(i_1,i_2)$, because each token points to its strictly best neighbour and the score is symmetric. So every cycle has length two.

Proof of Theorem 5.3. For two probability rows p, q, the identity $\min(p_k,q_k) = (p_k+q_k-|p_k-q_k|)/2$ proves the two formulas for Dobrushin's coefficient $\tau(P)$ (Dobrushin, 1956). Put $a_k = \min(p_k,q_k)$ and $t = 1-\sum_k a_k$. If $t = 0$ the rows agree. Otherwise $(p-a)/t$ and $(q-a)/t$ are probabilities; their weighted row vectors are two points of the convex hull of the rows of X. The diameter of that hull is at most $d(X)$: expand the difference of two convex combinations as a convex combination of pairwise differences and apply the triangle inequality. The difference of pX and qX is t times such a difference, proving the bound. For rows attaining the maximum, choose scalar $x_k = 1$ where $p_k > q_k$ and zero elsewhere. If $t > 0$, these inputs have diameter one and the two outputs differ by exactly t. If $t = 0$ the zero coefficient is already sharp. Finally apply the bound twice to PQX and use this sharp scalar characterization for PQ. □

Proof of Corollary 5.4. Apply Theorem 5.3, the operator norm of W_ℓ, and the same Lipschitz inequality for ϕ_ℓ to each pair of rows. A common bias cancels in that pairwise inequality. For the residual update, the triangle inequality bounds each row difference by $(1-p_\ell)d(X_\ell) + p_\ell d(F_\ell(X_\ell))$. Iterate and use $1-z \leq e^{-z}$. For sharpness use two scalar rows, identity features, and $P = \bar{q}I + (1-\bar{q})\mathbf{1}\mathbf{1}^\top/2$; every contrast is multiplied by exactly $1-p_\ell(1-\bar{q})$. □

Proof of Proposition 5.6. Diameter is subadditive, giving $d_{\ell+1} \leq a_\ell d_\ell + u_\ell$; induction gives the unrolled expression. A shared decoder makes every output pair distance at most $Gd(X_L)$, proving the second assertion. At exact consensus it maps identical inputs to identical outputs. □

A.7 Proofs for Chapter 6

Proof of Theorem 6.3. The projections U, Q_z and Q_y commute, so P_0 is diagonal on the orthogonal decomposition into constants, Z, the centered class space Y and interactions, with eigenvalues 1, a, b and 0. Since $\widehat{P}$ is symmetric and stochastic, the centered space is invariant under it, and by Weyl's inequality every centered eigenvalue moves by at most e. The top d_z centered eigenvalues of

$\widehat{P}$ are therefore at least $a-e$, and all others at most $b+e<a-e$. Let $\hat{v}$ be a unit eigenvector in the top band, with eigenvalue $\hat{\lambda}\geq a-e$. Then $(\hat{\lambda}I-P_0)\hat{v}=E\hat{v}$. The projection $I-\Pi_Z$ commutes with P_0, and on the centered part of $Z^{\perp}$ the operator $\hat{\lambda}I-P_0$ has eigenvalues $\hat{\lambda}-b$ and $\hat{\lambda}$, both at least $a-b-e>0$. Hence $\|(I-\Pi_Z)\hat{v}\|\leq\|E\hat{v}\|/(a-b-e)\leq e/(a-b-e)$. Summing the squares over an orthonormal basis of $\widehat{V}$ gives the Frobenius bound. For the resource bound, a leading centered eigenvalue is at most $a+e\leq Cw+e$, so $t\leq Cw+e$. For $E=0$ the filter multiplies the nuisance component by $g(a)$ and the class component by $g(b)$; the two are orthogonal, which gives the ratio. □

Proof of Theorem 6.4. From $\mathbb{E}e^{t\|z\|^2}=(1-2t)^{-d/2}$, Chernoff's inequality gives $\|z\|^2\leq d+2\sqrt{du}+2u$ outside probability e^{-u}, so $\|z\|\leq\sqrt{d}+\sqrt{2u}$ there; a union bound over the n points makes every noise vector at most R_δ in norm with probability at least $1-\delta$. On that event write a mapped center difference as u and its noise perturbation as v: $\|u\|\leq BD$, $\|v\|\leq 2\nu$ and $|\|u+v\|^2-\|u\|^2|\leq 4BD\nu+4\nu^2$. Exponentiating gives the weight bounds, and summing them gives the degree bounds. Each normalized entry then satisfies $e^{-2a'}(H_0)_{ij}\leq H_{ij}\leq e^{2a'}(H_0)_{ij}$, so $|H-H_0|\leq eH_0$ entrywise, and for real v, $|v^{\top}(H-H_0)v|\leq e\,|v|^{\top}H_0|v|\leq e\|v\|^2$, because H_0 is symmetric and similar to a stochastic matrix, with norm one. The variational principle gives both conclusions. The event does not depend on A or h, which proves uniformity. □

Proof of Proposition 6.5. The four-by-four affinity factors as the tensor product of the matrices $\left(\begin{smallmatrix}1 & k_j\\ k_j & 1\end{smallmatrix}\right)$ with $k_j=e^{-2w_jm_j^2/h^2}$, $j\in\{y,z\}$. Each has normalized eigenvalues 1 and $(1-k_j)/(1+k_j)=\tanh(w_jm_j^2/h^2)$, and the tensor product gives the four eigenpairs. Replicating the groups equally adds only zero eigenvalues. □

Proof of Proposition 6.6. (i) is linearity: $\sum_j P_{ij}Mk_j=M\sum_j P_{ij}k_j$, and a direction annihilated by W_K is annihilated by MW_K. If the query of a row is held fixed, changing a source token along $\ker W_K$ changes neither its key nor its value. (ii) Injective recovery forces E to be injective, so $r\geq\operatorname{rank}E=d_b$; at $r=d_b$, E is invertible and so is $E^{\top}$. (iii) $E^{\top}q=0$ for all queries in the span means that the range of E lies in the span's orthogonal complement, of dimension $r-s$, which holds an injective image of $\mathbb{R}^{d_b}$ exactly when $d_b\leq r-s$; any isometric embedding into it attains the bound. □

Proof of Proposition 6.7. $\|Eb\|\geq\|b\|/L_D$ for all b. Write $E=UC$ with $U^{\top}U=I$, so $C^2\succeq L_D^{-2}I$ and $\operatorname{tr}(E^{\top}GE)=\operatorname{tr}(C\,U^{\top}GU\,C)\geq L_D^{-2}\operatorname{tr}(U^{\top}GU)$. The last trace is smallest when U spans the d_b lowest eigenvectors of G. The choice $E=U/L_D$, $D=L_DU^{\top}$ attains equality. □

For a single perturbed source in fixed softmax rows the local metric is $G=\tau^{-2}\sum_i\omega_ip_{ij}(1-p_{ij})q_iq_i^{\top}$, since the Hessian of the log-partition gives $\sum_i\omega_iD_{\mathrm{KL}}(p_i\,\|\,p_i^e)=e^{\top}Ge/2+o(\|e\|^2)$. Width, weak query directions and decoder gain are therefore distinct resources.

Proof of Theorem 6.8. For each competitor j, $q^{\top}(k_*-k_j)=q^{\top}(a_*-a_j)+(E^{\top}q)^{\top}(b_*-b_j)+q^{\top}(\xi_*-\xi_j)\geq\Delta-2Be_q-2Q\nu=m$ by Cauchy–Schwarz. Hence $p_j/p_*\leq e^{-m/\tau}$ and $p_*=\big(1+\sum_{j\neq *}p_j/p_*\big)^{-1}\geq\big(1+(n-1)e^{-m/\tau}\big)^{-1}$. For the value error, expand $\sum_j p_jv_j-b_*=\sum_{j\neq *}p_j(b_j-b_*)+\sum_j p_j(v_j-b_j)$ and use the triangle inequality. The share bound is at least $1-\eta$ exactly when $m\geq\tau\log((n-1)(1-\eta)/\eta)$, which is the budget. For necessity, take all address gaps equal to Δ and $\|q\|=Q$, the target content antiparallel to $E^{\top}q$ and every competing content parallel to it, both of norm B, and the target key noise antiparallel to q and every competing noise parallel to it, both of norm ν. Every score difference then equals m, and the share equals its bound. □

For tied decoding with $Da_j = 0$ and $DE = I$, take $v_j = Dk_j$ and $\nu_v = L_D\nu$; the address-annihilation and decoding conditions must be feasible together, which is what Proposition 6.6 decides. For centered Gaussian key noise independent of the query, with covariance at most $\sigma^2 I$, all $|q^\top \xi_j|$ are at most $\sigma Q\sqrt{2\log(2n/\delta)}$ with probability at least $1-\delta$, and this quantity replaces $Q\nu$.

Proof of Theorem 6.9. Completing the square gives the posterior covariance $C = (\tau_b^{-2} I + \sigma_k^{-2} E^\top E)^{-1}$ and mean $CE^\top k/\sigma_k^2$, which minimizes squared loss among all functions of k. The optimum for fixed E is $\operatorname{tr} C$, which is the displayed sum in the squared singular values p_j of E, with $d - k_0$ unobserved directions. For given singular values in decreasing order, the smallest exposure is $\sum_j g_j p_j$, attained by pairing the largest with the smallest g_j: the squared overlaps of the two eigenbases form a doubly stochastic matrix, and the minimum linear pairing is the oppositely ordered one. The query-blind case confines E to an $(r-s)$-dimensional subspace, and Jensen's inequality for the convex map $p \mapsto \sigma_k^2/(\sigma_k^2/\tau_b^2 + p)$ makes the equal allocation P/k optimal on the k available directions. □

Proof of Theorem 6.11. Condition on $G = g$ and on the target's class y. The target's unhalved score is $c + g_y$. Each other item's unhalved score is $Z_j + g_{Y_j}$ with an independent uniform class, so it has distribution function F_g, independently across items given g. All $N-1$ fall below $c + g_y$ with probability $F_g(c + g_y)^{N-1}$, and ties have probability zero. Averaging over the target's class and over G gives p_N. With no cargo every threshold is c. For unequal class probabilities the same steps give weighted sums. □

The curve decreases in N term by term, strictly where $0 < F_G(c + G_y) < 1$ with positive probability. Its half-success count is where it crosses $\frac{1}{2}$. The evaluation quoted in Section 6.3 integrates the exact conditional law with six independently scrambled Sobol rules (2^{15} points each); their spread is a diagnostic, not an error bar. Evaluated on the archived coarse grid, the same model gives a tied onset within 0.002 of the archive's. The inputs, seeds and hashes are in `data/separation_quadrature/` of the package of Bruce (2026e) (Appendix D).

Proof of Proposition 6.12. Each coordinate maximizes its own addend independently, so the global argmaxes are exactly the pairs of half-argmaxes, and every Cartesian top-k list, which keeps a top half-score on each side, contains one of them. A selection that is a function of the clean scores alone cannot change when the cargo does. Comparing each perturbed pair score with the winner's gives both conditions. □

Inclusion for random codebooks. Condition on the query halves, the target codewords and the target's half-scores t_A, t_B. Put $\rho_h = \|q_h\|_2/\sqrt{2}$, and let F_{64} be the distribution function of a spherical cosine in $\mathbb{R}^{64}$. The other $B-1$ codewords in half h score independently with distribution function $F_{64}(x/\rho_h)$, so the target is included in a top-k product shortlist with conditional probability $\prod_{h\in\{A,B\}} \Pr\{\mathrm{Bin}(B-1, 1-F_{64}(t_h/\rho_h)) \le k-1\}$, by counting the codewords above the target. Averaging over the query and the targets gives the unconditional probability. For imperfect queries the half-norms fluctuate, so setting $\rho_h = \frac{1}{2}$ throughout is an approximation, not the source's distribution.

Proof of Theorem 6.13. Condition successively on no departure so far: its probability is at least $(1-\epsilon)^k$. Ending outside implies at least one departure. Average over T. Bernoulli's inequality gives $1-(1-\epsilon)^k \le k\epsilon$; Taylor's theorem with second derivative bounded by $k(k-1)$ gives the remainder. □

Proof of Proposition 6.14. B, U are commuting orthogonal projections with $BU = U$. The spectral identity follows by multiplying these projections. They give $\|P_\epsilon^L\|_F^2 = 1 + (K-1)a^2$. Off-block entries equal $(1-a)/n$, and their number is $n^2(1-1/K)$, giving the numerator. Differentiate the formula at $a = 1$ using $a'(0) = -L$. □

Proof of Theorem 6.15. U, $B - U$, and $I - B$ are mutually orthogonal symmetric projections of ranks $1, K-1, n-K$, since $B^2 = B$, $U^2 = U$, and $BU = UB = U$. The matrix P_ϵ acts on them by $1, 1-\epsilon, 0$. Functional calculus, or evaluation on each of the three ranges, gives the first identity. Frobenius orthogonality gives its squared norm as the sum of squared coefficients times ranks. Since I and B have no off-block entries, the off-block part is $(h-a)\,\mathrm{off}(U)$. There are $n^2 - \sum_b n_b^2$ such entries, each $1/n$, giving its squared norm $(h-a)^2\xi$. Division proves the claim. □

Proof of Corollary 6.16. The path mixture has $g(1) = 1$, $g(0) = p_0$, and $g(1-\epsilon) = A$, so apply Theorem 6.15. Because $0 \le [1-(1-\epsilon)^T]/\epsilon \le T$, dominated convergence gives $A'(0+) = -\mathbb{E}T$. The numerator vanishes at zero, so only its derivative contributes and the denominator there is $\sqrt{K + (n-K)p_0^2}$. At $\epsilon = 1$ one has $A = p_0$, proving saturation. Divide the slope by that of $T \equiv 1$ for the ratio. □

A.8 Proofs for Chapter 7

Proof of the convergence assertion of Theorem 7.1, and of Corollary 7.2. For a row-stochastic matrix P, the triangle inequality gives

$$\|(PX)_i\|_2 \le \sum_j P_{ij}\|x_j\|_2 \le \|X\|_{\infty,2}.$$

More generally, $\|MX\|_{\infty,2} \le \|M\|_\infty \|X\|_{\infty,2}$ for any real matrix M. Hence $\|F_A(X)\|_{\infty,2} \le \alpha R + \beta R = R$ on B_R. Splitting

$$F_A(X) - F_A(Y) = \alpha\mathcal{P}(X)(X-Y) + \alpha(\mathcal{P}(X) - \mathcal{P}(Y))Y$$

gives

$$\|F_A(X) - F_A(Y)\|_{\infty,2} \le \alpha\|X-Y\|_{\infty,2} + \alpha L_P R\|X-Y\|_{\infty,2}. \tag{A.1}$$

The closed finite-dimensional ball is complete, so the contraction mapping theorem gives a unique fixed point and

$$\|X_t - X^*\|_{\infty,2} \le q^t\|X_0 - X^*\|_{\infty,2}.$$

If the builder is defined and row stochastic outside B_R as well, any fixed point anywhere must satisfy $\|X^*\|_{\infty,2} \le \alpha\|X^*\|_{\infty,2} + \beta\|A\|_{\infty,2}$ and thus lies in B_R. Theorem 7.1 needs neither a global builder extension nor a convergence assertion for initializations outside the ball.

For two anchors $A, A' \in B_R$, both maps have the same contraction upper factor. Their fixed-point difference satisfies

$$\|X^*(A) - X^*(A')\|_{\infty,2} \le q\|X^*(A) - X^*(A')\|_{\infty,2} + \beta\|A - A'\|_{\infty,2},$$

and therefore

$$\|X^*(A) - X^*(A')\|_{\infty,2} \le \frac{\beta}{m}\|A - A'\|_{\infty,2}. \tag{A.2}$$

The common domain is essential: a bound computed only on the nominal anchor's ball is not automatically valid for a perturbed anchor outside it. The residual of any candidate $Y \in B_R$ also gives

$$\|Y - X^*\|_{\infty,2} \le \frac{\|Y - F_A(Y)\|_{\infty,2}}{1-q}. \tag{A.3}$$

This bound becomes weak near $q = 1$. A small absolute residual alone is insufficient to resolve a central difference at a still smaller perturbation amplitude. □

A diameter bound on a convex invariant domain. Let $C \subset \mathbb{R}^d$ be compact and convex, with diameter D_C, and assume all anchor rows lie in C. Then C^N is invariant. If $M\mathbf{1} = 0$, write each row of M as the difference of its positive and negative parts. Both parts have mass $\|M_{i,:}\|_1/2$, and their normalized averages of rows of $Y \in C^N$ lie in C. Hence

$$\|MY\|_{\infty,2} \le \tfrac{1}{2} D_C \|M\|_\infty. \tag{A.4}$$

Applying this to $M = \mathcal{P}(X) - \mathcal{P}(Y)$ proves the same contraction result with $\kappa_{\mathrm{f}} = L_P D_C/2$, where L_P is required only on C^N. For anchor initialization one may take $C = \mathrm{conv}\{a_i\}$, whose diameter is $\max_{ij}\|a_i - a_j\|_2$. For response bounds between different anchors, C must contain both anchors; differentiating on a full neighborhood requires a common domain containing that neighborhood. The geometric factor $D_C/2$ is independent of the coordinate origin and is at most R when $C \subset B(0,R)$. Translating coordinates also translates the builder's argument; a dot-product builder itself need not be translation invariant. The antipodal sharpness construction has $D_C = 2$, so this refinement preserves its constant.

Proof of the sharpness assertion of Theorem 7.1. Write $E = \mathbf{1}\mathbf{1}^\top/2$ and $Q = I - E$, and take the anchor $A = (1,-1)^\top$, with $\|A\|_{\infty,2} = 1$. We give a constructive proof of the sharpness assertion on B_1. Fix $\alpha \in (0,1)$, $\beta = 1-\alpha$, and $\varepsilon > 0$. Choose

$$0 < \theta \le \min\left\{\frac{\alpha}{4(1+\alpha)}, \frac{1}{8}, \frac{\varepsilon}{6(1+\varepsilon)}\right\}, \quad a_1 = 1 - 2\theta, \quad a_2 = 1 - \theta.$$

Both points lie strictly between $\beta/(1+\alpha)$ and one. Define

$$G(a) = \frac{a-\beta}{\alpha a}, \qquad k = \frac{G(a_2) - G(a_1)}{a_2 - a_1} = \frac{\beta}{\alpha a_1 a_2},$$
$$l(a) = G(a_1) + k(a - a_1), \qquad g(a) = \mathrm{clip}_{[-1,1]}\, l(a).$$

The function G is strictly increasing and strictly concave for $a > 0$. Thus $g(a_i) = G(a_i) \in (-1,1)$ and a_i is a fixed point of $f(a) = \alpha g(a) a + \beta$. On the linear piece,

$$f(a) - a = \alpha k (a - a_1)(a - a_2).$$

Since $l(1) > G(1) = 1$, the clipped builder is identically one near $a = 1$; this is a third fixed point. The multipliers at the first two are

$$f'(a_1) = 1 - \frac{\beta\theta}{a_1 a_2}, \qquad f'(a_2) = 1 + \frac{\beta\theta}{a_1 a_2}.$$

The first lies in $(0,1)$ because $a_1 a_2 \ge (3/4)(7/8)$ and $\beta\theta \le 1/8$; the second is greater than one. The multiplier at one is $\alpha < 1$. Thus the outer two fixed points attract and the middle one repels. The

sum coordinate of the full two-token state evolves as $x_1^+ + x_2^+ = \alpha(x_1 + x_2)$, so both scalar attractors are also attractors in the full state space.

The builder $P(X) = E + g((x_1 - x_2)/2)Q$ is stochastic because $|g| \le 1$. Its exact Lipschitz constant from $\|\cdot\|_{\infty,2}$ to $\|\cdot\|_\infty$ is k: the contrast coordinate has norm one, and antisymmetric perturbations in the linear segment attain k. Also

$$a_1 a_2 = 1 - 3\theta + 2\theta^2 \ge 1 - \frac{\varepsilon}{2(1+\varepsilon)} > \frac{1}{1+\varepsilon}, \qquad \beta < \alpha L_P = \frac{\beta}{a_1 a_2} < (1+\varepsilon)\beta.$$

Thus $m = \beta - \alpha L_P \in [-\varepsilon\beta, 0)$.

To obtain a smooth builder, convolve the globally defined clipped affine function g with an even, nonnegative, compactly supported C^∞ mollifier of mass one. Choose its support narrower than the distance from $a_1, a_2, 1$ to the two clip corners. Convolution preserves the range $[-1, 1]$ and the Lipschitz upper bound k. It leaves affine and constant pieces unchanged near the three roots (the mollifier's first moment is zero), and therefore preserves the fixed points and multipliers exactly. The resulting $\mathcal{P}$ is smooth on the entire state space. This proves the sharpness asserted in Theorem 7.1 without needing a statement about every orbit or every equality-case builder.

For completeness, any two fixed points $0 < a_1 < a_2 \le 1$ in this family obey

$$\frac{g(a_2) - g(a_1)}{a_2 - a_1} = \frac{\beta}{\alpha a_1 a_2}, \qquad \alpha L_P \ge \frac{\beta}{a_1 a_2} \ge \beta.$$

This necessity within the family requires Lipschitz continuity, not differentiability. It is consistent with the global sufficient condition and says nothing about necessity for uniqueness in an arbitrary individual map.

For any $N \ge 2$ and $d \ge 1$, embed the first two tokens along a unit vector, let their 2×2 block depend only on that scalar contrast, and give the remaining tokens identity blocks and zero anchors. Orthogonal coordinates and additional tokens contract with factor at most α. The builder's Lipschitz bound and the two attracting equilibria are unchanged. Sharpness within the narrower tied-dot-product or Gaussian builder classes is a separate question. □

Proof of Proposition 7.4. Let $p = \mathrm{softmax}(u)$. For any vector v,

$$\|D\,\mathrm{softmax}(u)v\|_1 = \sum_i p_i |v_i - \mathbb{E}_p v| \le \tfrac{1}{2}\,\mathrm{osc}(v) \le \|v\|_\infty, \qquad \mathrm{osc}(v) = \max_i v_i - \min_i v_i.$$

Indeed, subtract the midpoint of the range of v; the derivative annihilates constant vectors. After scaling to $|v_i| \le 1$, Cauchy–Schwarz bounds the mean absolute deviation by $\sqrt{\mathrm{Var}_p(v)} \le 1$. Integrating along the line segment between two logits proves the Lipschitz bound. On B_R,

$$s|\langle x_i, x_j\rangle - \langle y_i, y_j\rangle| \le s(\|x_i - y_i\|_2\|x_j\|_2 + \|y_i\|_2\|x_j - y_j\|_2) \le 2sR\|X - Y\|_{\infty,2}.$$

Applying the rowwise softmax bound proves $L_P \le 2sR$. A fixed mask applies the same argument to the fixed set of allowed logits in each row; each set must be nonempty. Fixed biases cancel in logit differences.

On a convex set $C \subset B(0, R)$ with diameter D_C, the oscillation of the directional logits in row i is bounded by

$$s\big|\langle u_i, x_j - x_k\rangle + \langle x_i, u_j - u_k\rangle\big| \le s(D_C + 2R)\|U\|_{\infty,2}.$$

Thus $L_P \le s(D_C/2 + R)$ on C^N, which can be combined with (A.4).

For K2, the Gaussian builder whose bandwidth depends on the evolving cloud, let $d_{ij}(X) = \|x_i - x_j\|_2^2$ and $v(X) = h_{\min}^2 + c_h \sum_{i \neq j} d_{ij}(X)/(N(N-1))$. On B_R,

$$d_{ij}(X) \le 4R^2, \quad |d_{ij}(X) - d_{ij}(Y)| \le 8R\|X - Y\|_{\infty,2}, \quad |v(X) - v(Y)| \le 8c_h R\|X - Y\|_{\infty,2}.$$

Since $v \ge h_{\min}^2 > 0$, adding and subtracting $d_{ij}(Y)/v(X)$ gives

$$\left| \frac{s\, d_{ij}(X)}{2v(X)} - \frac{s\, d_{ij}(Y)}{2v(Y)} \right| \le s \left(\frac{4R}{h_{\min}^2} + \frac{16 c_h R^3}{h_{\min}^4} \right) \|X - Y\|_{\infty,2}.$$

Thus this expression is a valid K2 builder bound. The experiments use $h_{\min} = 0.1$, $c_h = 1$, and diagnostic ball radius 1.25; prepared anchors have row norm below one. The bound treats distance and bandwidth variations separately. Controlling their joint variation on a tighter invariant domain may improve it; we establish no optimal K2 threshold. Sampled derivatives describe local sensitivity but cannot certify the complete ball. □

Proof of Proposition 7.5. $Q(X)X - Q(Y)Y = Q(X)(X - Y) + (Q(X) - Q(Y))Y$, whose norm is at most $(r_{\max} + RL_Q)\|X - Y\|_{\infty,2}$; invariance follows from $\alpha r_{\max} R + \beta R \le R$. For a sub-stochastic row p with mass r, the logit Jacobian $\partial p/\partial s = \mathrm{diag}(p) - pp^\top$ has $\|\partial p/\partial s\|_{\infty \to 1} = r - \min_{u \in \{\pm 1\}^N} (p^\top u)^2 \le r$: for a sign vector u with $m = p^\top u$, $\|(\partial p/\partial s)u\|_1 = \sum_j p_j |u_j - m| = r - m^2$, and the maximum of a norm of a linear map over the cube is attained at a vertex. Each score changes by at most $2|s|R\|\mathrm{d}X\|_{\infty,2}$, so each row of Q changes by at most $2|s|R\, r_{\max}\|\mathrm{d}X\|_{\infty,2}$ in ℓ_1; integrating along the segment, which stays in the ball, gives L_Q. □

Proof of Proposition 7.7. Part (a). For a probability vector p the softmax Jacobian is $\partial p/\partial s = \mathrm{diag}(p) - pp^\top$. For $\|v\|_\infty \le 1$, $\|(\partial p/\partial s)v\|_1 = \mathbb{E}_p|v - \mathbb{E}_p v| \le \sqrt{\mathrm{Var}_p(v)} \le 1$, so softmax is 1-Lipschitz from ℓ_∞ to ℓ_1; two equal probabilities and $v = (1, -1)$ attain the bound. On the ball B_R, Cauchy–Schwarz gives $\max_{i,j} |\tau x_i^\top x_j - \tau y_i^\top y_j| \le 2\tau R\|X - Y\|_{\infty,2}$. Splitting $P(X)X - P(Y)Y = P(X)(X - Y) + (P(X) - P(Y))Y$ bounds the first term by $\|X - Y\|_{\infty,2}$ and the second by $2\tau R^2\|X - Y\|_{\infty,2}$, so $\mathrm{Lip}(F) \le \alpha(1 + 2\tau R^2)$. Each output row is βx_{0i} plus α times a convex combination of rows of norm at most R, so the ball is invariant when $R \ge R_0$. A contraction of a closed set into itself has a unique fixed point that attracts every start, by the summable geometric series of consecutive differences.

Part (b). On the ball B_{R_0} every logit difference in a row is at most $2\tau R_0^2$, so $P_{ij}(X) \ge e^{-2\tau R_0^2}/m_i \ge \ell_i$ for $j \in V_i$. The image of the ball lies in $\mathcal{K}_1 = \prod_i [\beta x_{0i} + \alpha \overline{B}(0, R_0)]$. For $X \in \mathcal{K}_1$ write $x_j = \beta x_{0j} + \alpha z_j$ with $\|z_j\| \le R_0$, and split each row's probabilities into the common ℓ_i and the remainder of mass $1 - m_i\ell_i$:

$$F_i(X) = c_i + \alpha\beta \sum_{j \in V_i} (P_{ij} - \ell_i) x_{0j} + \alpha^2 \sum_{j \in V_i} P_{ij} z_j.$$

The last two terms have norm at most $\alpha\beta(1 - m_i\ell_i)R_0 + \alpha^2 R_0 = r_i$, so $F(\mathcal{K}_1) \subseteq \mathcal{K}$. Also $\|c_i - \beta x_{0i}\| + r_i \le \alpha R_0$, so $\mathcal{K} \subseteq \mathcal{K}_1$. This gives invariance and absorption in two steps, and every fixed point in the ball lies in $\mathcal{K}$. Part (a)'s argument on the convex set $\mathcal{K}$, whose rows have norm at most R_{eff}, then gives the contraction, and absorption gives attraction from the whole ball.

Part (c). For X in the candidate ball, $\|F(X) - \widehat{X}\|_{\infty,2} \le qr + e \le r$, so the ball is invariant and part (a)'s contraction argument applies; the fixed point X_* satisfies $\|X_* - \widehat{X}\|_{\infty,2} \le q\|X_* - \widehat{X}\|_{\infty,2} + e$. For the refinement, a zero-sum signed vector a has $\|\sum_j a_j y_j\| \le (D/2)\sum_j |a_j|$, since its positive and negative parts are equal masses times two convex combinations of rows, at distance at most D. Applying this to a row of $P(X) - P(Y)$, whose ℓ_1 norm is at most $2\tau R\|X - Y\|_{\infty,2}$, replaces $2\tau R^2$ by τRD. □

Proof of Remark 7.6. The upper bound $L_P \le 2sR$, and with it $\kappa_{\mathrm{f}} \le 2sR^2$, is Proposition 7.4. For the sharpness of the constant 2, fix $R > 0$, put $s = (\log 2)/(2R^2)$, and take three scalar tokens, with no mask and no bias, in the states

$$X = (R, -R, -R), \qquad Y_h = (R - h, -R + h, -R + h), \qquad 0 < h < R.$$

Both lie in B_R, and $\|X - Y_h\|_{\infty,2} = h$. With $r = R - h$, the first row of $\mathcal{P}(Y_h)$ gives the first token the weight

$$p(r) = \frac{e^{sr^2}}{e^{sr^2} + 2e^{-sr^2}} = \frac{e^{2sr^2}}{e^{2sr^2} + 2}$$

and each of the other two the weight $(1 - p(r))/2$. Since $e^{2sR^2} = 2$, $p(R) = 1/2$; put $\delta_h = 1/2 - p(R - h) > 0$. The first row of $\mathcal{P}(X) - \mathcal{P}(Y_h)$ is then $(\delta_h, -\delta_h/2, -\delta_h/2)$, whose absolute sum is $2\delta_h$. Differentiating,

$$p'(R) = \frac{8sR\, e^{2sR^2}}{(e^{2sR^2} + 2)^2} = sR, \qquad \text{so} \qquad \lim_{h \downarrow 0} \frac{2\delta_h}{h} = 2sR.$$

Every Lipschitz constant of this builder on B_R is therefore at least $2sR$, so here $L_P = 2sR$ and $\kappa_{\mathrm{f}} = 2sR^2$. Multiplying by the radius loses nothing either: the first row of the feedback term $(\mathcal{P}(X) - \mathcal{P}(Y_h))Y_h$ is $2(R - h)\delta_h$, and $2(R - h)\delta_h/h \to 2sR^2$. So for every $C < 2$ the bound with CsR^2 in place of $2sR^2$ fails for small h, and 2 is the least constant that holds for every scale, radius, dimension and number of tokens. This is a statement about that universal constant, not an equality for every builder (a single token gives a constant builder), and it does not make $2\alpha sR^2 < 1 - \alpha$ necessary for convergence. Finally, $8sR^2 = 4 \cdot 2sR^2$. □

A.9 Proofs for Chapter 8

Proof of Proposition 8.1. Differentiating $p_j = e^{s_j}/(e^b + \sum_k e^{s_k})$ gives $\partial p_j/\partial s_k = p_j(\delta_{jk} - p_k)$, that is $\partial p/\partial s = \mathrm{diag}(p) - pp^\top$, which is also the Hessian of $\log(e^b + \sum_k e^{s_k})$. For $x \in \mathbb{R}^N$, $x^\top(\partial p/\partial s)x = \sum_j p_j x_j^2 - (\sum_j p_j x_j)^2$, and Cauchy–Schwarz with weights p_j gives $(\sum_j p_j x_j)^2 \le r \sum_j p_j x_j^2$, hence $x^\top(\partial p/\partial s)x \ge (1 - r)\sum_j p_j x_j^2 \ge (1 - r)\min_j p_j \|x\|^2$. Finally $p = r\,\mathrm{softmax}(s)$ with $r = \sum_j e^{s_j}/(e^b + \sum_j e^{s_j}) = \sigma(\mathrm{LSE}(s) - b)$. □

For a row-stochastic P with stationary law π, Jensen's inequality gives $(Pv)_i^2 \le \sum_j P_{ij} v_j^2$, and summing against π with $\pi^\top P = \pi^\top$ gives $\|Pv\|_\pi^2 \le \|v\|_\pi^2$, with equality on constants.

Proof of Proposition 8.2. $P_{ij} = \dfrac{\sum_k e^{s_{ik}}}{z_i + \sum_k e^{s_{ik}}} \cdot \dfrac{e^{s_{ij}}}{\sum_k e^{s_{ik}}}$, and the first factor is $\sigma(\mathrm{LSE}_i(s) - \log z_i)$. □

Proof of Theorem 8.3. Perron–Frobenius supplies positive eigenvectors, and a positive root because Q is irreducible and nonzero. Multiplying $Q\mathbf{1} \le \mathbf{1}$, strict in some row, by the positive left eigenvector gives $\rho\nu^\top\mathbf{1} < \nu^\top\mathbf{1}$, hence $\rho < 1$. The equation $Qh = \rho h$ gives unit row sums for P^h. Summing $\nu_i h_i P^h_{ij} = \nu_i Q_{ij} h_j/\rho$ over i proves stationarity of $\pi^h \propto \nu \circ h$. The ratio $\pi^h_j P^h_{ji}/\pi^h_i$ equals $\nu_j Q_{ji}/(\rho\nu_i)$, which gives the reversal. Jensen's inequality and stationarity give $\|P^h\|_{L^2(\pi^h)} = 1$, and the diagonal similarity gives $\|Q\|_w = \rho$. For $Q_{ij} = e^{s_{ij} - L_i}$ the forward row normalizer is $\sum_j e^{s_{ij}} h_j = e^{L_i}\rho h_i$ and the reverse normalizer is $\sum_j e^{s_{ji}}\nu_j e^{-L_j} = \rho\nu_i$, which give the two softmax forms. With structural zeros the same formulas hold on the support. □

Strict killing in some row is not enough without irreducibility: a disconnected stochastic block can keep eigenvalue one. Uniform survival at most $r < 1$ in every row gives $\|Q^t\|_\infty \le r^t$ directly.

Proof of Proposition 8.4. The Perron expansion $Q^t = \rho^t h \nu^\top + o(\rho^t)$, divided by $\mu^\top Q^t \mathbf{1}$, gives the endpoint limit. For a path $x_0, \dots, x_t$, telescoping the ratios of h gives $\prod_{j<t} Q_{x_j x_{j+1}} = \rho^t (h(x_0)/h(x_t)) \prod_{j<t} P^h_{x_j x_{j+1}}$, so the survival-conditioned path law has a density proportional to $h(X_0)/h(X_t)$ relative to the P^h path with the same initial law. This density is bounded above independently of t. The law of large numbers for the primitive chain P^h concentrates its occupation on π^h, and the bounded density transfers the concentration to the conditioned law. □

Primitivity rules out periodic oscillation of the endpoint. The causal, reducible case needs its own accessibility analysis: $Q = \mathrm{diag}(0.9, 0.5)$ started in state 2 stays there whenever it survives, so conditioning cannot move it to the state with the larger diagonal, and Jordan chains can add polynomial factors. For an absorbing chain with transient block Q from which every state reaches the target, take the longest of the finitely many shortest absorbing paths and the smallest of their positive probabilities: some finite m has $\|Q^m\|_\infty < 1$, hence $Q^t \to 0$ and all mass is absorbed.

Proof of Theorem 8.5, in full. Order the tokens so that the first is the sink. Each R_ℓ is stochastic with first row $e_1^\top$, so $T_L = R_L \cdots R_1$ has the block form $\left(\begin{smallmatrix} 1 & 0 \\ b_L & B_L \cdots B_1 \end{smallmatrix}\right)$ with $b_L = \mathbf{1} - B_L \cdots B_1 \mathbf{1}$. For a probability row u, $\|u - e_1^\top\|_1 = (1 - u_1) + \sum_{j \geq 2} u_j = 2 \sum_{j \geq 2} u_j$, and for row $i \geq 2$ of T_L the off-sink mass is $(B_L \cdots B_1 \mathbf{1})_i$, the probability that the backward walk has not reached the sink. Taking the maximum over rows gives (8.4).

For the envelopes (with $T \geq 2$ and $i, j \geq 2$ indexing the non-sink tokens), $(B_\ell \mathbf{1})_i = 1 - p_{\ell,i} + p_{\ell,i}(Q_\ell \mathbf{1})_i = 1 - p_{\ell,i} s_{\ell,i}$ because $Q_\ell \mathbf{1} = \mathbf{1} - s_\ell$. Since $B_\ell \geq 0$ entrywise, $B_\ell v \leq \|v\|_\infty B_\ell \mathbf{1} \leq \|v\|_\infty (1 - \min_i p_{\ell,i} s_{\ell,i}) \mathbf{1}$ for $v \geq 0$, and induction gives the upper bound. For the lower bound, $(B_\ell v)_i \geq (1 - p_{\ell,i}) v_i \geq (1 - \max_j p_{\ell,j}) v_i$, and induction from $v = \mathbf{1}$ gives $B_L \cdots B_1 \mathbf{1} \geq \prod_\ell (1 - \max_j p_{\ell,j}) \mathbf{1}$. If every $Q_\ell = (1 - \delta_\ell) I$ (self-or-sink layers) and each p_ℓ is constant across tokens, then $B_\ell = (1 - p_\ell \delta_\ell) I$ and the upper bound holds with equality.

For a single kernel A and an independent proper path law, row i of $f(A)$ is the law of X_K from i, whose off-sink mass is $\mathbb{P}_i(\tau > K) = \sum_k p_k (Q^k \mathbf{1})_i$, with τ the absorption time at the sink. A leak $s_i \geq \delta$ gives $Q\mathbf{1} \leq (1 - \delta)\mathbf{1}$ and $Q^k \mathbf{1} \leq (1 - \delta)^k \mathbf{1}$, with equality for $Q = (1 - \delta) I$. The maximum over rows stays outside the expectation; exchanging them changes the quantity. For substochastic attention rows, append a virtual absorbing token whose column receives the missing mass $1 - \sum_j A_{ij}$; the augmented kernel is stochastic and the identity applies with the virtual token as the sink. □

Row-dependent laws. If row i uses its own proper law K_i, independent of the walk, the same argument gives $\|F - \mathbf{1} e_1^\top\|_\infty = 2 \max_i \mathbb{P}_i(K_i < \tau)$ with $F_{ij} = \sum_k p_{i,k} (A^k)_{ij}$. Such filters, as in node-dependent smoothing (Zhang et al., 2021), keep the survival identity but generally lose the scalar spectral reduction, because F no longer commutes with A.

Proof of Proposition 8.6. Triangular products. If every content operator is lower-triangular, then $(M_L \cdots M_1)_{ii} = \prod_\ell (M_\ell)_{ii}$: every path from i back to i through lower-triangular matrices must stay at i, because an index that decreases cannot increase again.

Layerwise differences. Fix a content row with sink mass $s < 1$, self-attention d, off-diagonal content mass $1 - s - d \geq 0$, and share p. From (8.5),

$$F_{ii} = 1 - p + pd + ps, \qquad C_{ii} = \frac{1 - p + pd}{1 - ps}, \qquad X_{ii} = 1 - p + \frac{pd}{1 - s}.$$

Direct algebra gives

$$F_{ii} - C_{ii} = \frac{p^2 s (1 - s - d)}{1 - ps} \geq 0, \qquad C_{ii} - X_{ii} = \frac{p(1 - p)\, s(1 - s - d)}{(1 - ps)(1 - s)} \geq 0,$$

with positive denominators. Multiplying the layerwise inequalities along the diagonal proves the ordering for products. The ordering concerns each token's weight on itself; it is not an ordering of every entry, of mixing rates, or of task error.

The skip switch. Under C the probability of leaving token i is $1 - C_{ii} = [(1 - ps) - (1 - p + pd)]/(1 - ps) = p(1 - s - d)/(1 - ps) = \pi$. With p and d held fixed, its derivative in s is $-p(1 - p + pd)/(1 - ps)^2 \leq 0$, so more sink mass cannot increase mixing. Under F the leaving probability is $p(1 - s - d)$, also nonincreasing in s. If $\pi_{\ell,i} < 1$ for all ℓ, then $\prod_\ell(1 - \pi_{\ell,i})$ has a positive limit if and only if $\sum_\ell \pi_{\ell,i} < \infty$, by the argument for fixed sequences in the complements to Theorem 2.11 (Section A.3). □

The forced anchor. The first content token has no other content token to attend to, so under all three treatments its row is absorbing even when it is not absorbing in the original kernel. Averages of self-weight in Section 2.7 therefore exclude it together with the sink. Normalizing only the complete killed product, $\mathrm{diag}(B_L \cdots B_1\mathbf{1})^{-1}B_L \cdots B_1$, is a fourth statistic and generally differs from $C_L \cdots C_1$, which normalizes at every layer.

Proof of Proposition 8.7. Part (a). Dividing numerator and denominator by e^b gives $a = 1/(1 + Be^{-b})$, which is at least $\frac{1}{2}$ exactly when $e^b \geq B$. With $B = Nm(\tau)$ this is $\ln N \leq b - \ln m(\tau)$. Under $b = b_0 + \ln N$ the mean-field abstention is $e^{b_0}N/(e^{b_0}N + Nm(\tau)) = e^{b_0}/(e^{b_0} + m(\tau))$. On a present query the target's weight is $e^{\tau c}/(e^{\tau c} + Ne^{b_0} + Nm(\tau))$, which falls below $\frac{1}{2}$ once $\ln N > \tau c - \ln(e^{b_0} + m(\tau))$.

Part (b). The rule abstains when all N independent scores fall below θ, with probability $F(\theta)^N$. Solving $F(\theta)^N = \frac{1}{2}$ gives the onset. If $F(\theta) = 1$ it never fails, and if $F(\theta) < 1$ it fails eventually, although no normalizer is involved.

Part (c). Given the row, the M samples are independent, and each lands above θ with probability p_θ. □

A finite number of competitive walkers turns (a)'s decision into a binomial absorption event. It is not deterministically equivalent to the half-mass decision, but converges to it, away from equality, as the number of samples grows.

A.10 Proofs for Chapter 9

Every finite identity and bound in the proofs for Sections 9.1 and 9.2 below was also checked numerically on random float64 instances (Bruce, 2026c); the checks are falsification tests, not proofs. Logarithms are natural.

Proof of Proposition 9.2. The targets read $z_0 \geq A_*B_0$ and $z_1 \leq M/W_* - B_1$, and the sensitivity bound gives $z_1 \geq e^{-\epsilon}z_0 \geq e^{-\epsilon}A_*B_0$, which proves necessity. Conversely, $z_0 = A_*B_0$ and $z_1 = e^{-\epsilon}A_*B_0$ are positive with log-ratio exactly ϵ; they meet abstention with equality and meet retrieval exactly when the condition holds. With bounded scores, $B_0, B_1 \geq Ne^{-\tau}$ and $M \leq e^\tau$ turn the condition into $N \leq N_\epsilon$. Since $1/(1 + A_*) = 1 - a_*$, $N_0 = N_*$, and N_ϵ increases to $e^{2\tau}(1 - w_*)/w_* = N_*/(1 - a_*)$. Distractors at $-\tau$, the match at τ and these nulls meet both targets for every $N \leq N_\epsilon$, with retrieval exactly w_* at $N = N_\epsilon$. For equal backgrounds $B_0 = B_1 = B$ with $r = M/(W_*B)$, some finite ϵ works exactly when $r > 1$, and the least one is $\max\{0, \log(A_*/(r - 1))\}$. For the last statement, absent success is $Z \geq A_*B$ and present success is $Z \leq M/W_* - B < A_*B$, so the event $\{Z \geq A_*B\}$ has probability at least $1 - \alpha$ under absence and at most β under presence. □

The fixed-sink bounds of Section 9.1 follow from $B_N \geq Ne^{-\tau}$ and $e^{s_*} \leq e^{\tau}$: $a_N = e^b/(e^b+B_N) \leq \sigma(b+\tau-\log N)$ and $w_N \leq e^{\tau}/(e^{\tau}+Ne^{-\tau})$.

Proof of Proposition 9.3. The absent constraint gives $N\sigma(-\tau+c_N) \leq M$. For $N > M$ this gives $e^{c_N} \leq e^{\tau}M/(N-M)$, hence $p_* \leq e^{2\tau}M/(N-M+e^{2\tau}M)$, and $p_* \geq w_*$ rearranges to the bound. Sigmoid weights without a row-mass guarantee have no probabilistic abstention mass, and the proposition must not be applied to them under that name. □

The proof of Proposition 9.4 comes after that of Theorem 9.8, whose tempered background sum B_N it uses, together with the bounded-variable bound of the proof of Proposition 9.5.

Proof of Proposition 9.5. For a variable in $[l,u]$ the exponential-moment bound $\mathbb{E}e^{t(X-\mathbb{E}X)} \leq e^{t^2(u-l)^2/8}$ holds, since the second derivative of the log moment-generating function is a variance under an exponential tilt, at most $(u-l)^2/4$, and its value and first derivative vanish at zero. Markov's inequality on the product of the moments, optimized over t and union-bounded over the two tails, gives the concentration of B_N/N, and monotonicity of $z/(z+B)$ gives the interval. Jensen gives $\mathbb{E}a_N \geq a_{\mathrm{mf}}$. Taylor's theorem with $f''(B) = 2z/(z+B)^3$, $B \geq Nl$ and $\mathrm{Var}(B_N) = Nv$ gives the upper bound. With only a finite variance, Chebyshev with $t = \sqrt{v/(N\delta)}$ replaces the exponential bound. □

Proof of Theorem 9.6. After scaling, the background sum is at most $Ne^{\lambda s_0} = N_0^{\lambda}e^{\lambda s_0}$. Dividing by the null's $e^{\lambda b}$ gives $a_N \geq (1+v_a^{\lambda})^{-1}$, and dividing by the match's $e^{\lambda s_*}$ gives the retrieval bound. Powers of numbers in $[0,1]$ are non-increasing in λ, and tend to zero for numbers below one. These are sufficient conditions, not necessary ones at each finite context. □

Proof of Theorem 9.7. Divide the null's denominator by $e^{\lambda b}$ and use $\sum_j e^{\lambda s_j} = NM_{\lambda,N}^{\lambda}$ with $N = N_0^{\lambda}$. For $\lambda \geq 1$, Jensen applied to $x \mapsto x^{\lambda}$ gives $M_{\lambda,N} \geq M_{1,N}$, and each term is at most $e^{\lambda s_{\max}}$; the reciprocal is decreasing in its background argument. □

The identity is exact for the empirical scores. For a fixed law replicated to N keys the empirical power mean is that law's; for random scores the population power mean $M_q = (\mathbb{E}e^{qS})^{1/q}$ gives a population-sum prediction, not an exact finite-sample probability. The realized a_N need not be monotone in N, and the $b = s_0$ counterexample assumes a background exactly at s_0.

Proof of Theorem 9.8. Write $\lambda_N = \log N/\log N_0$ and $B_N = \sum_{j\leq N} e^{\lambda_N(S_j-b)}$. Almost surely every score is at most s_0, so $B_N \leq e^{\lambda_N(h-b)}$. For each $\delta > 0$, $p_\delta = \mathbb{P}(S > s_0-\delta) > 0$, and the strong law for its indicators puts at least $Np_\delta/2$ scores above $s_0-\delta$ eventually, so $B_N \geq (p_\delta/2)e^{\lambda_N(h-b-\delta)}$. Taking logarithms, dividing by λ_N and letting $\delta \downarrow 0$ along a countable sequence gives $\log B_N/\lambda_N \to h-b$ almost surely, which proves both strict cases. At $b = h$, $B_N = N^{-1}\sum_j e^{-\lambda_N(s_0-S_j)}$: the proportion of scores equal to s_0 tends to p_0 and bounds it below, and for any $\delta > 0$ it is bounded above by the proportion above $s_0-\delta$ plus $e^{-\lambda_N\delta}$, whose limit superior is at most $p_\delta \downarrow p_0$. Hence $B_N \to p_0$. The retrieval background term is the same normalized sum times $e^{\lambda_N(h-s_*)}$, which with $e^{\lambda_N(b-s_*)}$ gives every retrieval limit. For unbounded support use a threshold $T > b - \log N_0$ with $\mathbb{P}(S > T) > 0$ in the tail-count bound. □

Independence is used only for the convergence of the fixed-threshold frequencies, so the conclusions hold for any sequence with those limits and the same upper bound, such as a stationary ergodic one. Repeating one random score forever is a stationary, nonergodic counterexample, whose threshold is the realized score plus $\log N_0$.

Proof of Proposition 9.4. The bounded-variable bound gives $\mathbb{P}(|N^{-1}\sum_j U_{j,N} - \mu_N| > t) \le 2e^{-2Nt^2}$; use $B_N = v^{\lambda_N} N^{-1}\sum_j U_{j,N}$ and monotonicity. Convexity of $x \mapsto (1+v^{\lambda_N}x)^{-1}$ gives the expectation bound. □

Proof of Theorem 9.9. For $Y = s_0 - S \ge 0$, integration by parts gives $\mathbb{E}e^{-\lambda Y} = \lambda\int_0^\infty e^{-\lambda y}\mathbb{P}(Y \le y)\,dy$. Substitute $u = \lambda y$ and multiply by λ^a. On $y \le \eta$ the tail assumption bounds the integrand by a constant times $e^{-u}u^a$, and on $y > \eta$ the integral is at most $\lambda^a e^{-\lambda\eta}$, so dominated convergence gives $\mu_N \sim C\Gamma(a+1)\lambda_N^{-a}$. The concentration argument with $t_N = \sqrt{2\log N/N}$ has summable failure probabilities $2N^{-4}$, so by Borel–Cantelli the deviation is eventually at most t_N almost surely. Since $t_N = o(\lambda_N^{-a})$, the empirical mean is asymptotic to μ_N, which tends to zero because $a > 0$, and substitution gives the three assertions. □

Proof of Proposition 9.10. Part (a). With an optional null weight $z \ge 0$, $w_t \ge \frac{1}{2}$ reads $e^{\tau s_t} \ge z + \sum_{j\ne t} e^{\tau s_j} \ge e^{\tau s_j}$ for every $j \ne t$, and $\tau > 0$ gives $s_t \ge s_j$.

Part (b). Under the factorization the posterior probability that the target is j, given the scores s, is $\ell(s_j)/\sum_k \ell(s_k)$ wherever the denominator is positive, because $g(s)$ and the uniform prior cancel. A randomized rule that names j with probability $q_j(s)$ succeeds, given s, with probability $\sum_j q_j(s)\Pr(\text{target} = j \mid s) \le \max_j \Pr(\text{target} = j \mid s)$. Since ℓ is nondecreasing, a largest score attains that maximum. Averaging over s shows that no rule beats the argmax, and ties can be broken among the posterior maximizers without changing the value. For independent backgrounds with density f_0 and an independent target with density f_1, $p_t(s) = \prod_k f_0(s_k)\cdot f_1(s_t)/f_0(s_t)$, which is the factorization with $\ell = f_1/f_0$.

Part (c). Condition on the target's score. The target is the strict argmax when all $N-1$ independent background scores fall below c, which has probability $F(c)^{N-1}$; ties have probability zero. Solving $F(c)^{N-1} = \frac{1}{2}$ gives N_{ev}. □

Remark A.5 (Exchangeability is not enough). Put the target uniformly at one of three positions. Give the other two the same score B, uniform on $(0,1)$, and the target an independent score with density $2x$ on $(0,1)$. The backgrounds are exchangeable and the marginal likelihood ratio $2x$ is increasing, yet the unique score that differs from the other two identifies the target surely, while the argmax succeeds only with probability $\Pr(T > B) = \mathbb{E}\,T = \frac{2}{3}$. The factorization in (b) fails because the background scores are not independent.

For random keys, $F = F_D$ is the distribution function of the cosine between a fixed unit vector and a uniform one in $\mathbb{R}^D$ ($D \ge 2$). Its density is $(1-x^2)^{(D-3)/2}/B(\frac{1}{2}, \frac{D-1}{2})$ on $(-1,1)$, and the tail expansion of Proposition 9.12(b) gives the rate $C(c)$ used in Section 9.3.

Proof of Proposition 9.11. Part (a). The largest-score sample is the target exactly when no sample lands above the target and at least one lands on it. No sample lands above it with probability $(1-q)^M$, and none lands above it or on it with probability $(1-q-w)^M$; the difference is the success probability. With $q = 0$, $1-(1-w)^M \ge \frac{1}{2}$ is $w \ge 1 - 2^{-1/M}$.

Part (b). The number of target samples is $\text{Bin}(M, w)$. Its probability of exceeding $M/2$ increases strictly with w on $(0,1)$, by a coupling through uniform random variables. At $w = \frac{1}{2}$, symmetry gives probability $\frac{1}{2}$ for odd M and $\frac{1}{2}(1-\Pr\{\text{Bin}(M,\frac{1}{2}) = M/2\}) < \frac{1}{2}$ for even M, so the half-success threshold is $\frac{1}{2}$ for odd M and larger for even M.

Part (c). Along a path from the root to leaf j the factors $Z_{\text{child}}/Z_{\text{parent}}$ telescope to $e^{\tau s_j}/Z_{\text{root}}$, the flat softmax. For a product score $a_i + b_j$ the partition function factors, so the target's weight is the product of its two half-weights. Each half has $k-1$ other shortlisted entries with half-scores at most u, so its half-weight is at least $[1+(k-1)e^{-\tau(c/2-u)}]^{-1}$, with equality when the others all

score u. The product of the two bounds is at least $\frac{1}{2}$ exactly when $(k-1)e^{-\tau(c/2-u)} \le \sqrt{2}-1$, that is $\tau(c-2u) \ge 2\ln((1+\sqrt{2})(k-1))$. For $k=1$, inclusion alone gives the target all the mass. □

Remark A.6 (Soft at every level is not enough). Take two groups at the root, one holding two leaves and the other one leaf. Give the groups equal parent probabilities and every leaf the score zero. The leaves get probabilities $(\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$, while the flat softmax gives each $\frac{1}{3}$. Only parent weights proportional to the subtree partition functions reproduce the flat softmax.

The mean-field forms of Proposition 9.11. Replacing the background sum by $Nm(\tau)$ gives the target weight $e^{\tau c}/(e^{\tau c} + Nm(\tau))$, whose majority onset is $N^* = e^{\tau c}/m(\tau)$. With $q=0$ the select onset is $N_M^* = N^*(1/w_M - 1)$, and since $w_M \sim \ln 2/M$, the log of the total count $1 + N_M^*$ gains one nat per nat of $\ln M$. These follow by solving for N. For independent backgrounds with $m(2\tau) < \infty$, Chebyshev's inequality bounds the relative error of the replacement: $\Pr(|\sum_j e^{\tau z_j}/(Nm(\tau)) - 1| > u) \le (m(2\tau)/m(\tau)^2 - 1)/(Nu^2)$. Where this bound is large, the mean-field onset is not certified. For $z \sim \mathcal{N}(0, 1/D)$, completing the square gives the mean-field onset $\ln N^* = \tau c - \tau^2/(2D)$, which is largest at $\tau = cD$. It is the true onset only for $\tau < cD$: there the background sum freezes onto its largest terms, and beyond it the true onset keeps rising toward the evidence wall instead of falling (Lucibello & Mézard, 2024).

Proof of Proposition 9.12. Part (a). Fano's inequality (Cover & Thomas, 2006), with the actual error probability $p \le \varepsilon$, gives $H(J \mid Y) \le h(p) + p\ln(N-1) \le \ln 2 + \varepsilon \ln N$. So $I(J;Y) = \ln N - H(J \mid Y) \ge (1-\varepsilon)\ln N - \ln 2$.

Part (b). Suppose a background key has a cosine at least the target's. Then a selecting reader does not pick the target, and a router whose weights increase with the score gives that key at least the target's weight. The two weights sum to at most one, so the target's weight is at most one half. The background cosines are independent with distribution function F_d and the target's cosine is c, so the target is the strict argmax with probability $F_d(c)^{N-1}$. This is one half exactly at N_{ev}.

For the tail, a coordinate of a uniform unit vector in $\mathbb{R}^d$ has density $(1-x^2)^m/B(\frac{1}{2}, m+1)$ with $m = (d-3)/2$. Put $x = c + t$, $a = 2c/(1-c^2)$ and $b = (1+c^2)/(1-c^2)^2$. Then $1 - x^2 = (1-c^2)(1 - at - t^2/(1-c^2))$ and $m\ln(1 - at - t^2/(1-c^2)) = -m(at + bt^2 + O(t^3))$, so

$$1 - F_d(c) = \frac{(1-c^2)^m}{B(\frac{1}{2}, m+1)} \int_0^{1-c} e^{-m(at+bt^2+O(t^3))}\, dt = \frac{(1-c^2)^{m+1}}{2cm\, B(\frac{1}{2}, m+1)} \Big(1 - \frac{1+c^2}{2c^2 m} + O(m^{-2})\Big)$$

by Watson's lemma ($\int_0^\infty e^{-m(at+bt^2)}dt = \frac{1}{am}(1 - \frac{2b}{a^2 m} + O(m^{-2}))$; the integral beyond a fixed neighbourhood of $t=0$ is exponentially smaller). Next, $1/B(\frac{1}{2}, m+1) = \Gamma(m+\frac{3}{2})/(\Gamma(\frac{1}{2})\Gamma(m+1)) = \sqrt{(m+1)/\pi}\,(1 - \frac{1}{8(m+1)} + O(m^{-2}))$. Put $n = d-1$, so that $m+1 = n/2$ and $(1-c^2)^{m+1} = e^{-nC(c)}$, and take logarithms. The leading terms give $nC(c) + \ln(c\sqrt{2\pi n})$. The terms of order $1/n$ add up to $-2 + (1+c^2)/c^2 + \frac{1}{4} = c^{-2} - \frac{3}{4}$: the first comes from $1/(2cm) = 1/(c(n-2))$, the second from Watson's correction and the third from the gamma ratio.

Part (c). Every other logit is at least $s_t - R$ and $z \ge 0$, so $w_t \le e^{s_t}/(e^{s_t} + (N-1)e^{s_t - R}) = 1/(1 + (N-1)e^{-R})$. Then $w_t \ge w^*$ gives $(N-1)e^{-R} \le (1-w^*)/w^*$. A selecting reader returns the argmax, whatever the gap.

Part (d). Write $X = (X_1, \ldots, X_N)$ for the values and $\hat{X}_i$ for the answer to query i. Given M, the answer $\hat{X}_i$ depends on X only through M. So data processing and Fano's inequality give $H(X_i \mid M) \le H(X_i \mid \hat{X}_i) \le h(\varepsilon) + \varepsilon\ln(e^v - 1)$, since $h(p) + p\ln(e^v - 1)$ increases in the error probability p up to the chance level $1 - e^{-v}$. Then $H(X \mid M) \le \sum_i H(X_i \mid M)$, and $S \ge H(M) \ge I(X;M) = Nv - H(X \mid M)$. The weaker form uses $h(\varepsilon) \le \ln 2$ and $\ln(e^v - 1) < v$. For membership, each state m accepts a set $A(m)$ of keys. A store K in state m has $K \subseteq A(m)$ and

$|A(m) \setminus K| \le \varepsilon(U-N)$, so $|A(m)| \le N + \varepsilon(U-N)$. At most $\binom{\lfloor N+\varepsilon(U-N)\rfloor}{N}$ stores can therefore share a state, while all $\binom{U}{N}$ stores must be placed among the e^S states. For fixed N the ratio of the binomials tends to ε^{-N} as $U \to \infty$ (Carter et al., 1978).

Part (e). The answers to all pairs determine the order, so the machine recovers the uniformly random permutation π with probability at least $1-\varepsilon$. Fano's inequality gives $H(\pi \mid M) \le \ln 2 + \varepsilon \ln N!$, and so $S \ge H(M) \ge I(\pi; M) \ge (1-\varepsilon)\ln N! - \ln 2$. □

Proof of Corollary 9.14. Parts (b) to (e) of Proposition 9.12 and Proposition 10.6(a); for (v), $t(N, K+1)/\binom{N}{2}$ tends to $K/(K+1) < \phi$. For $c < 1$, N_{ev} is finite. For $\varepsilon < 1 - e^{-v}$ the per-item term in (d) is positive, since $v - h(\varepsilon) - \varepsilon \ln(e^v - 1)$ is the divergence from the uniform law of the law that puts mass $1-\varepsilon$ on the true value and spreads the rest evenly. □

A.11 Proofs for Chapter 10

Proof of Theorem 10.1. (a) For $N \ge 3$ the successor and predecessor of i_t are distinct targets, so $q_t \le 1 - p_t$. Then $\mathcal{A} = \sum_t \log p_t - \sum_t \log q_t \ge \sum_t \log p_t - \sum_t \log(1-p_t) = \sum_t \operatorname{logit}(p_t)$, with equality iff $q_t = 1 - p_t$ for all t. Reindexing gives $\sum_t \log q_t = \sum_t \log P_{i_{t+1} i_t}$, so $\mathcal{A}$ is the cycle affinity, equal to $\sum_t r_{i_t i_{t+1}}$ for softmax attention by Theorem 3.2(b). If P is reversible, Kolmogorov's criterion (Kolmogoroff, 1936) gives $\prod_t P_{i_t i_{t+1}} = \prod_t P_{i_{t+1} i_t}$, so $\mathcal{A} = 0$, and the geometric mean of p_t/q_t is exactly one. Then $\sum \operatorname{logit}(p_t) \le 0$ gives $\mathrm{GM}(p) \le \mathrm{GM}(1-p)$; with $\mathrm{GM}(p) + \mathrm{GM}(1-p) \le \mathrm{AM}(p) + \mathrm{AM}(1-p) = 1$ this yields $\mathrm{GM}(p) \le \frac{1}{2}$. If $p_t \ge 1 - \varepsilon$ for all t, then $N \operatorname{logit}(1-\varepsilon) \le \mathcal{A}$.

(b) If $P = P_\sigma$ is reversible with respect to a positive π, then $P_{i\sigma(i)} = 1$ and detailed balance give $\pi_i = \pi_{\sigma(i)} P_{\sigma(i)i}$, which forces $P_{\sigma(i)i} = 1$, that is, $\sigma(\sigma(i)) = i$. For composition, on $\mathbb{Z}_N$ the reflections $i \mapsto -i$ and $i \mapsto 1 - i$ are involutions, hence symmetric permutation matrices, and their composition is the unit shift.

(c) For $1 \le i \le N-2$, detailed balance on the edge $(i, i+1)$ gives $\pi_{i+1}/\pi_i = P_{i,i+1}/P_{i+1,i}$. Row $i+1 \le N-1$ puts at least $1-\varepsilon$ on $i+2 \ne i$, so $P_{i+1,i} \le \varepsilon$, and the ratio is at least $(1-\varepsilon)/\varepsilon$. (It is well defined, since $P_{i+1,i} = 0$ would force $\pi_i P_{i,i+1} = 0$.) Multiplying over i gives the bound. For the construction, $s_{ij} = -\beta(|i-j|-1)^2 + cj$ is symmetric plus a key bias, hence reversible by Theorem 3.2(b). In row i the successor beats the predecessor by $2c$, the token itself by $\beta + c$ and the token $i+2$ by $\beta - c$.

(d) Let R reverse token order and suppose every score is $s_{ij} = f(x_i, x_j) + b(|i-j|)$, with all other operations applied tokenwise. Then the scores of the reversed input are $s'_{ij} = s_{R(i)R(j)}$, because $|R(i) - R(j)| = |i-j|$. So $P' = RPR^\top$ at every layer and the network is equivariant, whatever f is and whether or not its heads circulate. For tied rotary attention, $s_{ij} = q_i^\top \mathcal{R}_{j-i} q_j$ with orthogonal rotations satisfying $\mathcal{R}_d^\top = \mathcal{R}_{-d}$. Then $s_{ji} = q_j^\top \mathcal{R}_{i-j} q_i = q_i^\top \mathcal{R}_{j-i} q_j = s_{ij}$, so the scores are symmetric and $J = 0$. Reversing positions, however, maps $\mathcal{R}_{j-i}$ to $\mathcal{R}_{i-j} \ne \mathcal{R}_{j-i}$, so the kernel is not conjugated by R.

(e) Entries of one row share a normalizer, so $\log(p_t/P_{i_t i_t}) = s_{i_t i_{t+1}} - s_{i_t i_t} = \lambda(\phi_{i_t}^\top \phi_{i_{t+1}} - \|\phi_{i_t}\|^2) + b_{i_{t+1}} - b_{i_t}$. Around the cycle the key-bias terms telescope, and $\sum_t \|\phi_{i_t}\|^2 = \frac{1}{2}\sum_t (\|\phi_{i_t}\|^2 + \|\phi_{i_{t+1}}\|^2)$, so the sum equals $-\frac{\lambda}{2}\sum_t \|\phi_{i_t} - \phi_{i_{t+1}}\|^2$, which is at most zero because $\lambda \ge 0$. The predecessor sum is the same expression traversed backwards. The tied rotary scores of part (d) are $(\mathcal{R}_i q_i)^\top (\mathcal{R}_j q_j)$, a Gram matrix of the rotated queries. For $s = \frac{1}{2}(QK^\top + KQ^\top) = XMX^\top$ with M symmetric, the same computation gives $-\frac{1}{2}\sum_t (x_{i_t} - x_{i_{t+1}})^\top M (x_{i_t} - x_{i_{t+1}})$, which is positive whenever the differences lie in a negative eigenspace of an indefinite M. □

Proof of Proposition 10.2. P_α is a convex combination of P and $P^\dagger$ for $|\alpha| \le 1$, both stochastic with law π. Its flow is $D_\pi P_\alpha = A + \frac{\alpha}{2} C$, so the reciprocal traffic is A and the current is αC. Cut flows are

constant because $\sum_{i\in U,j\notin U} C_{ij} = \sum_{i\in U,j} C_{ij} - \sum_{i,j\in U} C_{ij} = 0 - 0$. The Dirichlet form is constant by Proposition 3.3(a). Next, $P_{-a} = S - aJ = (S + aJ)^\dagger$ and $\widetilde{P}_{-a} = \widetilde{P}_a^\top$. Transposition preserves the characteristic polynomial and the singular values, and similarity preserves the characteristic polynomial. Entropy production is invariant under $F \mapsto F^\top$, and $H_\pi(M) = H(D_\pi M) - H(\pi)$ is invariant because the joint entropy of $D_\pi P_a$ equals that of its transpose. For the nonnegativity interval, $S_{ij} + \alpha J_{ij} \geq 0$ for all pairs iff $|\alpha| \leq S_{ij}/|J_{ij}| = (F_{ij} + F_{ji})/|F_{ij} - F_{ji}| = \coth(\frac{1}{2}|\log(F_{ij}/F_{ji})|)$ whenever $J_{ij} \neq 0$. □

Proof of Theorem 10.3. (a) Let R be reversible with respect to $\mu > 0$, so $W = D_\mu R$ is symmetric and $\log R_{ij} = \log W_{ij} - \log \mu_i$. Similarly $\log S_{ij} = \log A_{ij} - \log \pi_i$. Then

$$D(P\|R) - D(P\|S) - D(S\|R) = \sum_{ij}(F_{ij} - A_{ij}) \log \frac{S_{ij}}{R_{ij}} = \frac{1}{2}\sum_{ij} C_{ij}\Big[\log \frac{A_{ij}}{W_{ij}} - \log \frac{\pi_i}{\mu_i}\Big] = 0,$$

because the skew C annihilates the symmetric bracket and $C\mathbf{1} = 0$ annihilates the row term. Edges with $A_{ij} = 0$ carry no flow. Since $D(S\|R) \geq 0$ with equality iff $R = S$ (as $\pi > 0$), S is the unique minimizer. Along the path, P_α has law π and reciprocal part S, so the argument for Theorem 3.1 gives $D(P_\alpha\|S) = H_\pi(S) - H_\pi(P_\alpha)$. This is convex in α because KL is convex in its first argument and P_α is affine in α, and even because $P_{-\alpha} = P_\alpha^\dagger$.

(b) G is symmetric and nonnegative; by irreducibility (which holds for full-support softmax attention), its Perron root ρ and Perron vector v are positive. $E_{ij} = G_{ij}v_j/(\rho v_i)$ has row sums $(Gv)_i/(\rho v_i) = 1$. With $w = v^2/\|v\|^2$, $w_i E_{ij} = v_i v_j G_{ij}/(\rho\|v\|^2)$ is symmetric, so E is reversible with stationary law w. Now write any reversible R as $D_\mu^{-1}W$ with $W = D_\mu R$ symmetric and $\mathbf{1}^\top W\mathbf{1} = 1$, where μ is its stationary law. If $W_{ij} > 0$ where $G_{ij} = 0$, then $E_{ij} = 0$ and $P_{ij}P_{ji} = 0$, so $D(R\|P) = D(R\|E) = \infty$; otherwise W is supported where $P_{ij}, P_{ji}, E_{ij} > 0$, $D(R\|P) - D(R\|E) = \sum_{ij} W_{ij}\log(E_{ij}/P_{ij})$, and

$$\log \frac{E_{ij}}{P_{ij}} = \frac{1}{2}\log \frac{P_{ji}}{P_{ij}} + \log v_j - \log v_i - \log \rho.$$

The first term is antisymmetric and the next two form a potential difference; both vanish when summed against the symmetric W. The constant contributes $-\log\rho$ since $\sum W = 1$. Hence $D(R\|P) = D(R\|E) - \log\rho$ for every reversible R. Taking $R = E$ gives $D(E\|P) = -\log\rho$, so $D(R\|P) = D(R\|E) + D(E\|P)$. Because $D(R\|E) \geq 0$ with equality iff $R = E$, E is the unique minimizer, and $-\log\rho = D(E\|P) \geq 0$ gives $\rho \leq 1$, with equality iff $E = P$, that is, iff P is reversible. For softmax attention, $G_{ij} = e^{(s_{ij}+s_{ji})/2}/\sqrt{z_i z_j}$, and row-normalizing $G_{ij}v_j$ gives $\mathrm{softmax}(\frac{1}{2}(s + s^\top) + \mathbf{1}\beta^\top)$ with $\beta_j = \log v_j - \frac{1}{2}\log z_j$. □

Proof of Proposition 10.4. $(P^\dagger)_{ji} = \pi_i P_{ij}/\pi_j = \exp(\lambda q_i \cdot k_j + \log\pi_i - \ell_i)/\pi_j$. As a function of i for fixed j, this is proportional to $\exp(\lambda k_j \cdot q_i + b_i)$. The normalizer is $\sum_i \pi_i P_{ij} = (\pi^\top P)_j$, which equals π_j exactly when π is stationary; in general $\ell'_j - \log\pi_j = \log((\pi^\top P)_j/\pi_j) = \log(P^\dagger\mathbf{1})_j$. The update $\pi \leftarrow \pi^\top P$ is power iteration, and ℓ' computes it in log space. □

Proof of Proposition 10.5. (a) By the fundamental theorem of calculus applied to $t \mapsto \ell(Z_S + tB_J)$, which is exact because Z is affine in α before downstream recomputation. Taylor's theorem about 0 with $|\ell'''| \leq M_3$ gives $|o(a) - a\ell'(0)| \leq M_3 a^3/6$. The Simpson bound is the classical composite-rule remainder on $[-a, a]$ with one panel.

(b) With $p(\alpha) = \mathrm{softmax}(z(Z_S + \alpha B_J))$ we have $p'(0) = H_0 g$. Expanding both KL terms of the JS divergence about the mean of the two arms gives $\mathrm{JS} = \frac{a^2}{2}\sum_k (p'_k)^2/p_{0k} + O(a^4)$, with no

cubic term because JS is even in a. Moreover $\sum_k (H_0 g)_k^2 / p_{0k} = g^\top H_0 D_{p_0}^{-1} H_0 g = g^\top H_0 g$, since $H_0 D_{p_0}^{-1} H_0 = H_0$. The remainder requires four bounded derivatives near Z_S and probabilities bounded away from zero.

(c) $\pi^\top P_\alpha = \pi^\top S + \alpha \pi^\top J = \pi^\top$ because $\pi^\top J = \frac{1}{2}(\pi^\top P - \pi^\top P^\dagger) = 0$. The class-token row of $P^\dagger$ is (3.1) with $i = \text{cls}$. □

Proof of Proposition 10.6. A partial order. For positions $j < k$ write $j \prec k$ when $b_k - b_j \geq \ln r$, so that the ordered pairs are exactly the pairs related by $\prec$. The relation is transitive, since $j \prec k \prec l$ gives $b_l - b_j \geq 2 \ln r$, so it is a strict partial order. A chain $j_0 \prec j_1 \prec \cdots \prec j_m$ has $b_{j_m} - b_{j_0} \geq m \ln r$, so $m \leq B / \ln r$ and every chain has at most $K + 1$ elements.

Levels. Let $\ell(j) + 1$ be the number of elements of the longest chain that ends at j, so that $\ell(j) \in \{0, \ldots, K\}$. If $j \prec k$, appending k to a longest chain ending at j gives $\ell(k) \geq \ell(j) + 1$. Two positions on the same level are therefore never ordered with each other: the levels are Mirsky's partition (Mirsky, 1971) into antichains. Along a longest chain ending at j the positional term rises by at least $\ln r$ per step, so $b_j - \min b \geq \ell(j) \ln r$. Replacing b by $b'_j = \ell(j) \ln r$ therefore keeps every ordered pair, because $\ell(k) - \ell(j) \geq 1$ whenever $j \prec k$, and lowers every positional term relative to its minimum.

Part (a). Ordered pairs join different levels. If n_k positions sit on level k, there are at most $\frac{1}{2}(n^2 - \sum_k n_k^2)$ of them, the number of pairs split by a partition into at most $K + 1$ parts. Over integer parts this is largest when the parts are as equal as possible, where it equals $t(n, K+1)$. By Cauchy–Schwarz, $\sum_k n_k^2 \geq n^2/(K+1)$, so $t(n, K+1) \leq \frac{K}{K+1} \frac{n^2}{2}$ and $t(n, K+1) / \binom{n}{2} \to K/(K+1)$. The staircase has range $K \ln r \leq B$. Positions in different blocks differ by at least $\ln r$ in the right direction and positions in the same block not at all, so it orders exactly the $t(n, K+1)$ cross-block pairs. Ordering a fraction ϕ in the limit requires $K/(K+1) \geq \phi$, that is $\lfloor B / \ln r \rfloor \geq \phi/(1-\phi)$.

Part (b). Two independent uniform positions are ordered when they differ by at least $x = \ln r / \lambda$, which has probability $(1-x)_+^2$. The unordered fraction $1 - (1-x)^2 = 2x - x^2$ is about $2 \ln r / B$ at large B, while the staircase leaves $1/(K+1) \approx \ln r / B$.

Part (c). A single target at the position j_0 with the smallest positional term has weight $e^{\tau c_t + b_{j_0}}$ against the background sum $\sum_j e^{\tau c_j + b_j}$, which is $m(\tau) \sum_j e^{b_j}$ in mean field. Keeping half the mass is therefore $\sum_j e^{b_j - b_{j_0}} \leq e^{\tau c_t - \ln m(\tau)} = N e^{\Delta}$. The constraint only gets easier when positional terms fall relative to their minimum, so by the level argument the optimum is attained by a level assignment $b_j = \ell_j \ln r$ with $\sum_j r^{\ell_j} \leq N e^{\Delta}$. For given level counts n_k, sorting the levels by position orders every cross-level pair, and no arrangement orders a same-level pair, so the count is $\frac{1}{2}(N^2 - \sum_k n_k^2)$. With $f_k = n_k / N$ the ordered fraction tends to $1 - \sum_k f_k^2$, and the constraint reads $\sum_k f_k r^k \leq e^{\Delta}$. This is a strictly convex quadratic program, and its KKT conditions give $f_k = (\mu - \nu r^k)_+$ with $\nu \geq 0$. If $\Delta = 0$, then $\sum_k f_k r^k \leq 1 = \sum_k f_k$ with $r^k > 1$ for $k \geq 1$ forces $f_0 = 1$ and $F^* = 0$. For finite Δ, the levels above k_0 carry mass at most $e^{\Delta} r^{-k_0}$. Choosing k_0 with $e^{\Delta} r^{-k_0} \leq \frac{1}{2}$ leaves mass at least $\frac{1}{2}$ on $k_0 + 1$ levels, so $\sum_k f_k^2 \geq \frac{1}{4}/(k_0 + 1) > 0$ and $F^* < 1$.

Part (d). A timestamp comparison selects the later write, and overwrite leaves only one. Neither adds a positional term to the scores, so nothing is taken from the retrieval margin. □

Proof of Theorem 10.8. Pairing (i, j) with (j, i) gives the second form of σ, which equals

$$\sum_{i,j} \pi_i J_{ij} \ln[(A_{ij} + \pi_i J_{ij}) / (A_{ij} - \pi_i J_{ij})].$$

For $|x| < 1$, integrating the geometric series of $2/(1-x^2)$ gives $\ln[(1+x)/(1-x)] = 2 \sum_{k \geq 0} x^{2k+1}/(2k+1)$. Substituting $x = \pi_i J_{ij} / A_{ij}$ gives the series, each term of which is nonnegative and even. For

$k \geq 1$, $1/(2k+1) \leq \frac{1}{3}$ and successive powers shrink by at most η^2, which bounds the remainder. Edges with zero flow contribute nothing. □

A.12 Proofs for Chapter 11

Proof of Proposition 11.1. Given the realized input, the product forms convex combinations of its rows, and its rank is at most that of its right factor; iterating gives both statements. For the writer, the rows $(1, -1)$ and $(-1, 1)$ form a matrix of rank one, and the coordinatewise ReLU maps it to the identity, of rank two. □

A.13 Proofs for Chapter 12

Proof of Proposition 12.1. For fixed row-stochastic P, $\|\alpha P\|_\infty = \alpha < 1$. The Neumann series (Section A.1, T4) gives

$$R = \beta(I - \alpha P)^{-1} = \sum_{k \geq 0} \beta \alpha^k P^k.$$

Each P^k is stochastic and the scalar weights sum to one. Therefore R is stochastic, nonnegative, and fixes every constant-row state. With $P = P_*$ this proves (12.1) and frozen nonexpansiveness. Its induced maximum-row norm is exactly one (attained by constant rows). Its Euclidean norm need not be one. If $Pv = \mu v$, the modal gain is $\beta/(1 - \alpha\mu)$, whose magnitude is at most one for $|\mu| \leq 1$. This eigenvector statement does not bound the Euclidean singular values of a nonnormal map. For real $\mu \in [0, 1]$, the modal gain lies between β and one.

Define $\Psi(X, A) = X - \alpha\mathcal{P}(X)X - \beta A$. If $\mathcal{P}$ is C^1 near (X^*, A) and $D_X\Psi = I - DF$ is invertible, the implicit function theorem gives a differentiable branch and

$$(I - DF)D_A X^* = \beta I.$$

The chain rule yields the two terms in DF from (12.2). Positivity of $\mathcal{P}$ constrains its values but not the signed matrix $D\mathcal{P}[U]$; the latter has zero row sums. The inverse response may therefore have negative entries. If $\rho(DF) < 1$, standard local linear stability gives attraction in a suitable equivalent norm; the induced norm in the chosen coordinates may still exceed one. Under a valid differentiable common-ball contraction bound, $\|DF\| \leq q$ and the Neumann estimate gives $\|H\| \leq \beta/(1 - q)$ in the maximum-row norm, consistent with (A.2). □

A value bound against the nominal-frozen endpoint. There is also a useful perturbation bound relative to the nominal-frozen endpoint $X^0 = \beta(I - \alpha\mathcal{P}(A))^{-1}A$, the output of weights built once at the input. Every adaptive fixed point in B_R satisfies

$$\|X^* - X^0\|_{\infty,2} \leq \frac{2\alpha^2\kappa_{\mathrm{f}}R}{\beta}.$$

Subtract the two fixed-point equations, absorb the frozen mixing difference with factor α, and use $\|X^* - A\|_{\infty,2} \leq 2\alpha R$. This is a value bound; it is not a derivative bound and does not imply class preservation when initial margins are small. On C^N, both endpoints lie in the same convex row domain and $\|X^* - A\|_{\infty,2} \leq \alpha D_C$. Applying (A.4) instead gives the tighter value bound

$$\|X^* - X^0\|_{\infty,2} \leq \frac{\alpha^2 L_P D_C^2}{2\beta} = \frac{\alpha^2\kappa_{\mathrm{f}}D_C}{\beta},$$

with the hull constant $\kappa_{\mathrm{f}} = L_P D_C/2$ of Section A.8.

Proof of Proposition 12.2. Tightness at a prescribed contraction margin. Fix $0 < \alpha < 1$, $\beta = 1 - \alpha$, and $0 < m < \beta$. Set $k = (\beta - m)/\alpha$. For $0 < a < 1$ and $0 < \mu < 1$, put

$$A = a(1, -1)^\top, \quad z_* = \frac{\beta a}{1 - \alpha\mu}, \quad g(z) = c_w(\mu + k(z - z_*)), \quad \mathcal{P}(X) = E + g((x_1 - x_2)/2)Q,$$

where $E = \mathbf{1}\mathbf{1}^\top/2$, $Q = I - E$, and c_w is a smooth convolution of the clipping map to $[0, 1]$ with an even probability mollifier of support $[-w, w]$. Choose $w < \min(\mu, 1 - \mu)$, so c_w is the identity near μ. This builder is stochastic, $L_P = k$, and its update is contractive on the unit ball with upper factor $\alpha(1 + k) = 1 - m$. Since $a < 1$, the same ball contains a full neighborhood of the anchor. At $X^* = z_*(1, -1)^\top$, direct differentiation gives

$$R_* = E + \frac{\beta}{1 - \alpha\mu}Q, \qquad H = E + \frac{\beta}{1 - \alpha\mu - \alpha k z_*}Q. \tag{A.5}$$

Letting $a \uparrow 1$ and $\mu \uparrow 1$ sends R_* to I and the contrast response to β/m. For parameters sufficiently close to these limits that response exceeds one, so each cross-token derivative is negative. The maximum-row and Euclidean operator norms both equal the contrast response there. At $m = \beta$, the constant builder $P = I$ attains the bound one exactly. Consequently the response bound is optimal at every admissible positive contraction margin, including those bounded away from zero.

An explicit diverging family. Write $E = \mathbf{1}\mathbf{1}^\top/2$ and $Q = I - E$. Fix $0 < \varepsilon < 1/2$, $\alpha = \beta = 1/2$, $A = (1, -1)^\top$, $\mu = 1 - \varepsilon^2$, $k = 1 - \varepsilon$, and $z^* = 1/(1 + \varepsilon^2)$. Begin with the scalar function

$$c(u) = \operatorname{clip}_{[0,1]} u, \qquad c_w(u) = \frac{1}{2w}\int_{-w}^{w} c(u + t)\, dt, \qquad w = \varepsilon^2/4.$$

This box smoothing is C^1, stays in $[0, 1]$, is 1-Lipschitz, and equals u on $[w, 1 - w]$. It is the smoothing used by the experiment. A compact even C^∞ mollifier of the same support gives a C^∞ variant with the same local formulas. Set

$$g_\varepsilon(z) = c_w(\mu + k(z - z^*)), \qquad \mathcal{P}(X) = E + g_\varepsilon((x_1 - x_2)/2)Q.$$

The builder is stochastic and $L_P = k$. Since $\mu \in (w, 1 - w)$, it is exactly affine near z^*. The equilibrium equation is

$$z^* = \tfrac{1}{2}\mu z^* + \tfrac{1}{2}, \qquad X^* = z^* A.$$

On the unit ball, the contraction upper factor is $(1 + k)/2 = 1 - \varepsilon/2 < 1$. To justify differentiating with respect to a full neighborhood of the boundary anchor, use the larger radius $R_\varepsilon = 1 + \varepsilon/2$. The same globally Lipschitz builder gives

$$q_{R_\varepsilon} = \tfrac{1}{2}(1 + kR_\varepsilon) = 1 - \varepsilon/4 - \varepsilon^2/4 < 1.$$

All anchors of row norm at most R_ε share this invariant contraction ball.

The consensus direction has Jacobian eigenvalue $1/2$ and the contrast direction has eigenvalue $j = (\mu + kz^*)/2$. Consequently

$$H = E + G_\varepsilon Q, \qquad G_\varepsilon = \frac{1 + \varepsilon^2}{\varepsilon + 2\varepsilon^2 + \varepsilon^4}, \qquad R_* = E + \frac{1}{1 + \varepsilon^2}Q. \tag{A.6}$$

Together with the fixed-margin family, this proves Proposition 12.2. □

In this two-dimensional symmetric example, both the maximum-row and Euclidean induced norms of H equal G_ε when $G_\varepsilon \geq 1$. Each finite-ε response is bounded; the asymptotic divergence $G_\varepsilon \sim 1/\varepsilon$ accompanies a vanishing contraction margin and a shrinking admissible linear neighborhood. Higher-dimensional embeddings use anchor rows $(v, -v)$ for a unit vector v and define $z = \langle x_1 - x_2, v\rangle/2$; the displayed response is exact on the informative two-token subspace. State perturbations orthogonal to v do not change the builder and contract with factor at most α.

Proof of Proposition 12.4. Let R_* be the frozen smoother of (12.1). For a fixed affine readout,

$$X^*W + \mathbf{1}b^\top = R_*AW + R_*\mathbf{1}b^\top = R_*(AW + \mathbf{1}b^\top).$$

This proves score transport and, by separating its diagonal contribution, (12.10). □

For any candidate endpoint X with residual $r = X - \alpha PX - \beta A$ at its realized frozen P, and with $R = \beta(I - \alpha P)^{-1}$,

$$X - RA = (I - \alpha P)^{-1}r, \qquad \|X - RA\|_{\infty,2} \leq \|r\|_{\infty,2}/\beta.$$

This frozen reconstruction estimate needs no global contraction of the adaptive builder. It explains why even an uncertified settled endpoint can support a tightly checked value decomposition.

The uniform smoother. If $U = \mathbf{1}\mathbf{1}^\top/N$, then $U^2 = U$ and $R_U = \beta I + \alpha U$. Therefore the centered component is multiplied by β, its squared Frobenius dispersion by β^2, and each score retains self-weight $\beta + \alpha/N$.

Proof of Corollary 12.5. Positivity alone does not preserve class margins: even if the own-image term in (12.10) is positive, its companions can provide a larger negative term. A simple sufficient condition follows by setting $b_{ik} = \max_{j\neq i}[-(Z_{0,jy} - Z_{0,jk})]_+$. The entries of R_* are nonnegative and its off-diagonal entries in row i sum to $1 - (R_*)_{ii}$, so the companion contribution in (12.10) is at least $-(1 - (R_*)_{ii})b_{ik}$, and the margin remains positive whenever

$$(R_*)_{ii}(Z_{0,iy} - Z_{0,ik}) > \big(1 - (R_*)_{ii}\big)b_{ik}$$

for every competing class k. This is a task-dependent condition beyond stochasticity or convergence. □

Proof of Proposition 12.6. For symmetric nonnegative affinities $K = K^\top$ with positive row sums d_i, the kernel $P_{ij} = K_{ij}/d_i$ and $\pi_i = d_i/\sum_j d_j$ satisfy

$$\pi_i P_{ij} = K_{ij}\Big/\sum_j d_j = \pi_j P_{ji},$$

and summing over i shows that π is stationary. For fixed P, $\|\alpha P\|_\infty = \alpha < 1$, so $I - \alpha P$ is invertible (Section A.1, T4), and so is $\beta(I - \alpha P)^{-1}$. □

Proof of the margin identity at every iteration, with its affine-bias term (Section 12.5). Let R be any real mixing matrix, with row masses $\rho_i = \sum_j R_{ij}$, and let the readout be affine, with weights W and bias b, so that the anchors score $Z_0 = AW + \mathbf{1}b^\top$. The mixed state RA then scores

$$Z = RAW + \mathbf{1}b^\top = RZ_0 + (\mathbf{1} - R\mathbf{1})b^\top,$$

and for a true class y and a rival class c,

$$Z_{iy} - Z_{ic} = R_{ii}(Z_{0,iy} - Z_{0,ic}) + \sum_{j \neq i} R_{ij}(Z_{0,jy} - Z_{0,jc}) + (1 - \rho_i)(b_y - b_c).$$

These are the own and companion terms of (12.10), for R in place of R_*, and the affine-bias term, which vanishes when the row mass is one.

Every iterate of the anchored update is such a mixing of the anchors. Let $X_{t+1} = \alpha Q(X_t)X_t + \beta A$ with $X_0 = A$ and $\alpha + \beta = 1$, where Q is the softmax builder $\mathcal{P}$ of (1.4) or nonnegative weights whose rows may lose mass, as in a head that can abstain. Then $X_t = R_t A$ with $R_0 = I$ and $R_{t+1} = \alpha Q(X_t)R_t + \beta I$, by induction, since $X_{t+1} = (\alpha Q(X_t)R_t + \beta I)A$; so the identity holds at every iteration with $R = R_t$. At a fixed point X^* with $Q_* = Q(X^*)$, and the spectral radius of αQ_* below one, rearranging $X^* = \alpha Q_* X^* + \beta A$ gives $X^* = \beta(I - \alpha Q_*)^{-1}A$, which is (12.1) when $Q_* = P_*$. No derivative of Q enters: these are identities at the realized weights, not formulas for the response to the anchors. If the rows of Q sum to one, the same induction gives $R_t\mathbf{1} = \mathbf{1}$, so every row mass is one and the bias term vanishes. If rows of Q lose mass, it gives $R_t\mathbf{1} \leq \mathbf{1}$ entrywise, and the term $(1 - \rho_i)(b_y - b_c)$ enters wherever a row of R_t has lost mass. □

A.14 Proofs for Chapter 13

Proof of Corollary 13.1. Let C_δ project each row onto the Euclidean ball of radius δ, as in the centred outer map (13.2). That map fixes A_0, with inactive clipping in a neighbourhood of that anchor. The chain rule proves (13.1). Iterating it gives, with products ordered along the trajectory and an empty product equal to the identity,

$$H_K = DF_{K-1} \cdots DF_0 H_0 + \beta \sum_{t=0}^{K-1} DF_{K-1} \cdots DF_{t+1}.$$

This formula applies to a fixed iteration budget. It does not differentiate a discontinuous stopping-time decision. If $\|DF_t\| \leq q < 1$ in one fixed norm along the entire trajectory, then

$$\|H_K\| \leq q^K \|H_0\| + \beta \frac{1 - q^K}{1 - q} =: b_K. \tag{A.7}$$

For anchor initialization, $H_0 = I$. Under the global assumptions of Theorem 7.1, the same Lipschitz estimate holds for the finite map even without differentiability. Thus, on a domain where the outer map is invariant, a sufficient outer contraction condition is

$$(1 - \zeta) + \zeta\gamma b_K < 1, \qquad 0 < \zeta \leq 1, \quad \gamma \geq 0,$$

because rowwise projection onto a Euclidean ball is nonexpansive in the maximum-row norm. Locally at A_0, clipping is the identity and differentiation gives (13.3).

Now suppose $\|H_0\| \leq 1$. Since $q = 1 - m$ and $0 < m \leq \beta$, (A.7) is a convex combination of $\|H_0\|$ and β/m, which is (13.4), so

$$b_K \leq \frac{\beta}{m} \quad \text{for all } K \geq 0.$$

Consequently $\gamma < m/\beta$ suffices for outer contraction for every budget, for input initialization and for any fixed detached initializer, on a common invariant domain. □

The rowwise clipping in (13.2) supplies an invariant outer neighbourhood when its anchor rows, and all inner initializers, are contained in the common domain. A detached initializer need not equal the reference equilibrium for the bound; the experiment of Section 13.5 uses input initialization. That experiment does not claim this global sufficient condition for its focal example with the Gaussian builder of adaptive bandwidth: it computes the actual finite derivative and observes a local instability outside that guarantee.

Proof of the random-probe reading of a rank-one gain (Section 13.6). Let $H \in \mathbb{R}^{n\times n}$ be a fixed linear response, such as an outer loop's derivative, and let the probe u be uniformly distributed on the unit sphere of $\mathbb{R}^n$. Flipping the sign of one coordinate, or swapping two, leaves the law of u unchanged, so $\mathbb{E}u_iu_j = 0$ for $i \neq j$ and all the $\mathbb{E}u_i^2$ are equal; since $\sum_i u_i^2 = 1$, $\mathbb{E}uu^\top = I/n$. Hence

$$\mathbb{E}\|Hu\|_2^2 = \operatorname{tr}(H^\top H\, \mathbb{E}uu^\top) = \|H\|_F^2/n.$$

For a rank-one gain $H = Gvv^\top$ with $\|v\|_2 = 1$ and $G \geq 0$, $\|H\|_F = \|H\|_2 = G$, so the probe's root-mean-square response is $G/\sqrt{n}$, while a probe along v reads the full gain G. The identity is an average over directions: it shows how a directional probe can under-read an expanding mode, not how far any particular probe falls short. □

Proof of the two-loop certificate (Section 13.8). Let the joint update T act on states x made of one block x_i per loop, $i = 1, \dots, N$, in a convex domain, and let its valid upper gain bounds take the comparison form

$$\|T_i(x) - T_i(y)\|_i \leq \sum_j B_{ij}\|x_j - y_j\|_j, \qquad B = I - D_\beta + K,$$

with $D_\beta = \operatorname{diag}(\beta_i)$, $0 < \beta_i \leq 1$, $K \geq 0$ and $B \geq 0$ entrywise; uniform bounds on the block derivatives give such bounds. Write $M = D_\beta - K$.

The general certificate. For a nonnegative matrix A, $\rho(A) < 1$ holds exactly when some vector $v > 0$ has $Av < v$ entrywise. If $\rho(A) < 1$, the series $v = \sum_{k\geq 0} A^k\mathbf{1}$ converges by Gelfand's formula (Section A.1, T19), is at least $\mathbf{1}$, and satisfies $Av = v - \mathbf{1} < v$. Conversely, if $Av < v$, the weighted maximum norm $\max_i |z_i|/v_i$ gives A the induced norm $\max_i (Av)_i/v_i < 1$, so $\rho(A) < 1$. For $A = B$ and for $A = D_\beta^{-1}K$ the condition $Av < v$ reads $Kv < D_\beta v$, that is $Mv > 0$. Hence $\rho(D_\beta^{-1}K) < 1$, $\rho(B) < 1$ and the existence of $v > 0$ with $Mv > 0$ are equivalent. Given such a v, put $q = \max_i (Bv)_i/v_i < 1$ and $\|x\|_v = \max_i \|x_i\|_i/v_i$. The comparison bound gives

$$\frac{\|T_i(x) - T_i(y)\|_i}{v_i} \leq \sum_j \frac{B_{ij}v_j}{v_i}\,\frac{\|x_j - y_j\|_j}{v_j} \leq q\|x - y\|_v,$$

so T contracts in $\|\cdot\|_v$, and on a closed invariant domain it has a unique fixed point that attracts every start geometrically. The bounds alone certify nothing more: the linear map $T(x) = Bx$ on scalar blocks satisfies every one of them, and when $\rho(B) \geq 1$ its Perron eigenvector, nonnegative with eigenvalue $\rho(B)$ (Seneta, 1981), does not decay.

Two loops. Put $m_i = \beta_i - K_{ii}$ and $H = \begin{pmatrix} m_1 & -\sqrt{K_{12}K_{21}} \\ -\sqrt{K_{12}K_{21}} & m_2 \end{pmatrix}$. If $m_1, m_2 > 0$ and $m_1m_2 > K_{12}K_{21}$, the vector $v = (m_2 + K_{12},\, m_1 + K_{21}) > 0$ satisfies $Mv = (m_1m_2 - K_{12}K_{21})\mathbf{1} > 0$. Conversely, $Mv > 0$ with $v > 0$ gives $m_1v_1 > K_{12}v_2 \geq 0$ and $m_2v_2 > K_{21}v_1 \geq 0$, so $m_1, m_2 > 0$, and multiplying the two inequalities gives $m_1m_2 > K_{12}K_{21}$. These conditions say exactly that $H \succ 0$, so a positive pencil certifies the coupled pair, one-way coupling included. In the construction of Section 13.8, $m_i = 0.2$ and $K_{12} = K_{21} = g$, and the pencil is positive exactly when $g < 0.2$.

More loops. For N loops let H have diagonal entries m_i and off-diagonal entries $-\sqrt{K_{ij}K_{ji}}$. If the gains are balanced, $h_iK_{ij} = h_jK_{ji}$ for some $h_i > 0$, then $S = \operatorname{diag}(\sqrt{h_i})$ makes SKS^{-1} symmetric, with off-diagonal entries $\sqrt{K_{ij}K_{ji}}$, so $SBS^{-1} = I - H$. Since $B \geq 0$, $\rho(B)$ is its largest real eigenvalue (Seneta, 1981), which is $1 - \lambda_{\min}(H)$; so $\rho(B) < 1$ exactly when $H \succ 0$. Without balance the pencil can certify an unstable system. Let C be the cyclic permutation of three coordinates, $\beta_i = 1$ and $K = 1.1C + 0.01C^\top$. Then $B = K$ has constant row sums, so $\rho(B) = 1.11$, while H has unit diagonal and off-diagonal entries $-\sqrt{0.011}$, so $\lambda_{\min}(H) = 1 - 2\sqrt{0.011} > 0.79$.

Signs. The comparison sees only the sizes of the couplings, so its failure does not show that a particular signed system is unstable. With opposing couplings, $DF = \left(\begin{smallmatrix} 0.8 & 0.3 \\ -0.3 & 0.8 \end{smallmatrix}\right)$ fails the two-loop test, since $0.2^2 < 0.3^2$, yet contracts: its eigenvalues $0.8 \pm 0.3i$ have modulus $\sqrt{0.73} < 1$. □

A.15 Proofs for Chapter 14

Proof of Proposition 14.1. Write $\beta_0 = 1 - \mu_0 > \beta$, $r = \beta/\beta_0$ and $h(x) = -\ln(1 - x)$. Depth L gives $\Sigma = e^{Lh(\beta_0)}$ and $\kappa_{\text{col}} = e^{Lh(\beta)} = \Sigma^q$ with $q = h(\beta)/h(\beta_0)$, which is $\ln(1/\mu_C)/\ln(1/\mu_0)$; a constant-share stack, $f(\mu) = (1 - b + b\mu)^L$, gives $\kappa_{\text{col}} = \Sigma^{q_b}$ with $q_b = h(b\beta)/h(b\beta_0)$; the residual stream, $f(\mu) = e^{-\sigma(1-\mu)}$, gives $\Sigma = e^{\sigma\beta_0}$ and $\kappa_{\text{col}} = e^{\sigma\beta} = \Sigma^r$; and the restart, $1/f(\mu) = 1 + t(1 - \mu)$ with $t = \alpha/(1 - \alpha)$, gives $\Sigma = 1 + t\beta_0$ and $\kappa_{\text{col}} = 1 + t\beta = 1 + r(\Sigma - 1)$. The elasticity $xh'(x)/h(x)$ of h is increasing on $(0, 1)$, because its logarithmic derivative $1/x + 1/(1 - x) - h'(x)/h(x)$ is nonnegative exactly when $h(x) \geq x$, which holds on $(0, 1)$; hence $h(ry)/h(y)$ decreases in y, and $q = q_1 \leq q_b \leq \lim_{b\to 0} q_b = r$. Finally $\Sigma^r \leq 1 + r(\Sigma - 1)$, because $x \mapsto x^r$ is concave and $1 + r(x - 1)$ is its tangent at $x = 1$. □

A.16 Proofs for Chapter 15

Proof of Theorem 15.1. Necessity follows by applying the putative compensator. Conversely define $c(h(x)) = f(x)$; constancy on fibres makes this well-defined. The displayed inequality is exactly its Lipschitz condition and also forces fibre constancy. In the linear case define $C(Ax) = Bx$ on the image and extend it linearly to a complement. Kernel inclusion is precisely well-definedness. Taking orthogonal complements gives the row-space equivalence. □

Proof of Theorem 15.2. Necessity and the lower bound follow by row sums and the triangle inequality. Write row positive and negative masses p_i, n_i, so $p_i - n_i = s$. With $m = \|T\|_\infty$, set $a = (m + s)/2$, $b = (m - s)/2$. Then $\delta_i = a - p_i = b - n_i \geq 0$. Add the same nonnegative probability row with mass δ_i to both positive and negative parts. The resulting matrices have constant row masses a, b and difference T. Divide by their positive row masses to obtain $T = aP_+ - bP_-$; omit a zero-mass term. The cost is $a + b = m$. A scalar multiple of one positive head has only one sign. For strictly positive heads use $m' > m$ and positive slack: this gives cost arbitrarily close to m. □

A.17 Proofs for Chapter 16

Proof of Proposition 16.1. Write F for the feasible set. Every $t \in F$ has $0 \leq t_a \leq T/\beta_{n,a}$, because the rates are positive and the strengths nonnegative, so F is bounded even where a cap is infinite. It is therefore a nonempty bounded polyhedron: the convex hull of its finitely many vertices (Rockafellar, 1970, §§18–19).

A vertex attains the minimum. Write any $t \in F$ as a convex combination $t = \sum_v \lambda_v v$ of vertices. Jensen's inequality for the concave C (Section A.1, T6) gives $C(t) \geq \sum_v \lambda_v C(v) \geq \min_v C(v)$, so the

minimum of C on F exists and is attained at a vertex. No continuity is needed, since the inequality holds at every feasible point, not only at a minimizer.

At most one coordinate between its bounds. If two coordinates $a \neq b$ of some $t \in F$ lie strictly between their bounds, then $t \pm \varepsilon(e_a/\beta_{n,a} - e_b/\beta_{n,b})$ keeps the budget and, for small $\varepsilon > 0$, the caps. So t is the midpoint of two distinct feasible points and is not a vertex.

Strict concavity. If t minimizes C but is not a vertex, then $t = (y+z)/2$ for distinct $y, z \in F$, and strict concavity gives $C(t) > (C(y)+C(z))/2 \geq \min(C(y), C(z))$, which contradicts minimality. So every minimizer is a vertex. □

A.18 Proofs for Chapter 17

Chapter 17 defers no proof to this appendix: it applies the theorems of Chapter 9, whose proofs are in Section A.10.

A.19 Proofs for Chapter 18

Proof of Proposition 18.1. The final state is the ordered product applied to the start. Commutativity regroups it as $\prod_a A_a^{n_a} x_0$, independent of the order, with n_a the count of the symbol a, and a fixed readout preserves equality. □

Proof of Proposition 18.2. Row i of $(I - \alpha P)Y = \beta X$ reads $(1 - \alpha p_{ii})y_i - \alpha \sum_{j<i} p_{ij} y_j = \beta x_i$; solve for y_i. Substochasticity gives $p_{ii} \leq 1$, so $1 - \alpha p_{ii} \geq 1 - \alpha = \beta$ and the recursion exists and is unique. Under the factorization the prefix sum is $\sum_{j<i} p_{ij} y_j = u_i^\top H_{i-1}/z_i$ with $H_{i-1} = \sum_{j<i} v_j y_j^\top$, whose update is the outer-product update of the statement; multiplying numerator and denominator by z_i gives the second form. □

Proof of Proposition 18.3. The write is $M \mapsto MA + \eta v k^\top$ with $A = I - \eta(\lambda I + k k^\top)$. The symmetric matrix A has the eigenvalue $1 - \eta(\lambda + \|k\|^2)$ along k and $1 - \eta\lambda$ on the orthogonal complement of k, so $\|A\|_2$ is at most the larger of their absolute values (in one dimension only the first is present). Both eigenvalues are at most one, and $\eta\lambda \leq \eta(\lambda + \|k\|^2) \leq 2$ keeps both at least -1, so $\|A\|_2 \leq 1$. Two states written with the same key and value differ by $(M - M')A$, and $\|(M - M')A\| \leq \|M - M'\| \, \|A\|_2 \leq \|M - M'\|$, in the Frobenius norm and in the operator norm alike. This proves (a).

For (b), the function $x \mapsto |1 - \eta(\lambda + x)|$ is convex, so on $[0, K^2]$ it is at most the larger of its two endpoint values, which is q; hence $\|A_t\|_2 \leq q$ for every write. Since $\|v_t k_t^\top\| = \|v_t\| \, \|k_t\| \leq VK$ in either norm, $\|M_{t+1}\| \leq q\|M_t\| + \eta VK$, and summing the geometric series gives $\|M_t\| \leq q^t \|M_0\| + \eta VK(1 - q^t)/(1 - q)$. □

Proof that a non-expansive write can grow like $\sqrt{t}$ (Section 18.2). Take $\eta = 1$, $\lambda = 0$, two key dimensions and one value coordinate, so that the state is a row vector x and the write is $x \mapsto x + (v - \langle x, k \rangle)k$. Let $x_n = \big((-1)^n, \sqrt{n+1}\big)$ and $d_n = x_{n+1} - x_n$, and write with the unit key $k_n = (-1)^{n+1} d_n / \|d_n\|_2$ and the value $v_n = \langle k_n, x_{n+1} \rangle$. Since k_n is a unit vector along d_n, the write sends x_n to $x_n + \langle k_n, d_n \rangle k_n = x_n + d_n = x_{n+1}$. Each write is non-expansive, as $\eta(\lambda + \|k_n\|^2) = 1 \leq 2$, yet $\|x_n\|_2 = \sqrt{n+2}$ grows without bound. The keys' directions converge: $d_n = \big(2(-1)^{n+1}, \delta_n\big)$ with $\delta_n = \sqrt{n+2} - \sqrt{n+1} \to 0$, so $k_n = \big(2, (-1)^{n+1}\delta_n\big)/\|d_n\|_2 \to (1, 0)$. The values conflict: since $\delta_n(\sqrt{n+2} + \sqrt{n+1}) = 1$, $\langle d_n, x_{n+1} \rangle = 2 + \delta_n \sqrt{n+2} = (\|d_n\|_2^2 + 1)/2$, so $v_n = (-1)^{n+1}(\|d_n\|_2 + \|d_n\|_2^{-1})/2$ alternates in sign, and $5/4 \leq |v_n| \leq 3/\sqrt{5}$ because $2 \leq \|d_n\|_2 \leq \sqrt{5}$. Decay rules this out by part (b); without it, $q \geq |1 - \eta\lambda| = 1$ for every step. □

Proof of Proposition 18.4. Collect the keys as the columns of $K_0 \in \mathbb{R}^{d_k \times N}$ and the values as the columns of V_0, so that a linear memory M retrieves every pair exactly when $MK_0 = V_0$. If K_0 has rank N, then $K_0^\top K_0$ is invertible, and $M = V_0(K_0^\top K_0)^{-1}K_0^\top$ retrieves every assignment. If its rank is below N, some $c \neq 0$ has $K_0 c = 0$; exact retrieval would force $V_0 c = MK_0 c = 0$, which fails for the assignment $V_0 = wc^\top$ with any $w \neq 0$. A matrix of rank N with d_k rows has $N \leq d_k$, so one head stores at most d_k pairs exactly. □

A.20 Proofs for Chapter 19

Chapter 19 defers no proof to this appendix: its two propositions are proved where they are stated, and the entropy-production theorem that Section 19.3 recalls is proved in Section A.11.

A.21 Proofs for Chapter 20

Chapter 20 states predictions and the tests that will score them; it defers no proof to this appendix.

A.22 Proofs for Chapter 21

Chapter 21 defers no proof to this appendix: the design rules it collects follow from results of Chapter 2 and Section 8.3, which are proved where they are stated or in Sections A.3 and A.9, and the rest of the chapter gathers limits and open problems.

Appendix B

Evidence ledger

Table B.1 lists every claim of Chapter 0 and the decisive tests, with their labels, the sections of this book that develop each and the records behind it. It is generated from the public fields of the claims ledger behind Chapter 0, the source of Tables 0.1 to 0.4 and of the register of open tests (Table 20.1); the claims export in Chapter 0's package (`data/claims`, Appendix D) holds each record in full, with its plain and technical statements, its evidence and its caveats.

Table B.1: **The claims ledger.** Every claim of Chapter 0 and its decisive tests, grouped by the section of Chapter 0 that makes each, with its evidence label (Section R.2), its transfer label and that label's note (Section 0.5), the sections of this book that develop it, and the records behind it, named as in each paper's package (Appendix D), with their tier where the claim carries one. The usual answers carry the transfer label literature: explanations from the literature, not tested by the collection, whose bearings are in Table 0.1. The decisive tests carry the label pending: each result will be labelled when it returns (Section 20.4). A dagger marks an exact identity: it holds by construction for every operator of its class, and its value is what it makes measurable.

Claim	Evidence	In trained networks (transfer label and note)	Sections	Records
The usual answers (Section 0.2)				
It fit the hardware: large dense matrix multiplications	conventional	literature	14.4	—
Scaling laws made capability predictable and purchasable	conventional	literature	21.4	—
The ecosystem co-evolved around it	conventional	literature	14.2	—
Proving an alternative at scale costs too much	conventional	literature	15.6 and 18.5	—
History: the right objective, data and hardware in one place	conventional	literature	16.5	—
We measure what transformers are good at	conventional	literature	3.5	—
In-context learning turned one model into a platform	conventional	literature	21.4	—
Weak assumptions plus abundant data win	conventional	literature	1.1	—
Training is parallel, and every position gives a signal	conventional	literature	18.4	—

Claim	Evidence	In trained networks (transfer label and note)	Sections	Records
Exact, content-addressable memory makes recall possible	conventional	literature	9.6 and 18.3	—
The residual stream is a bus; blocks communicate, then compute	conventional	literature	5.1 and 11.7	—
It trains: conditioning is half the design	conventional	literature	2.6, 5.3 and 7.5	—
Data-dependent routing built from matrix multiplications	conventional	literature	1.3 and 14.4	—
Softmax can sharpen into a pointer	conventional	literature	9.1 and 9.4	—
The architecture is not the algorithm	conventional	literature	15.1	—
It absorbs its challengers as components	conventional	literature	16.5 and 19.6	—
Challengers converge towards attention	conventional	literature	18.1 and 19.6	—
Weaknesses get worked around above the architecture	conventional	literature	18.1 and 21.4	—
The design space around it is flat	conventional	literature	15.4	—
Scale sorts architectures, so far in the transformer's favour	conventional	literature	15.6	—
The transformer endures as a template for learned computation	conventional	literature	14.5	—
A sweet spot of conditions (Section 0.3)				
A mixing layer is an averaging operator rebuilt from its input	proved	exact: Every softmax head, at every layer and input, is a row-stochastic kernel $P(X)$ whose outputs are convex combinations of its values. The path-lottery form $\mathbb{E}[P^K]$ describes powers of one operator; a trained stack is an ordered product of different $P(X)$, with MLPs between.	1.1 to 1.4	—
The same laws govern graphs, denoisers and attention	proved	idealized: The collapse law is stated for powers of one fixed operator, which linear propagation over a fixed graph satisfies exactly; the certificate is stated for one builder $P(\cdot)$ applied repeatedly, as in iterative denoisers, deep equilibrium models and test-time mixing. A transformer's forward pass applies a different $P(X)$ at each layer.	4.1, 4.2 and 4.4	E3.2, N30 (SYN-R, REAL)

Claim	Evidence	In trained networks (transfer label and note)	Sections	Records
Deep averaging forgets unless some paths stay short	proved	idealized: The if-and-only-if is proved for path laws over powers of one fixed operator. Bounds for realized products of distinct layers and a Dobrushin bound for sink-like attention carry its direction, that mixing without short paths forgets, to stacks whose operators change from layer to layer.	2.2, 2.3, 2.6, 5.1 to 5.4 and 5.6	N25, E04, N37, SX43 (REAL)
Routing and content need separate maps	proved	exact under premises: Exact on the identified block model where necessity is stated; whether a trained network sits in that model must be checked, and with spare capacity trained networks evade it (compensation).	6.1 to 6.3 and 6.6	N12b, N11, E2.2 (REAL)
Self-reinforcing weights need their feedback held below the anchor	proved	exact under premises: The per-layer bound $\kappa_f \le 2sR^2$ (at logit scale s and state radius R) holds for every softmax layer with tied dot-product scores (projected scores multiply it by the projections' norms). Contraction and a single attracting state are statements about one builder applied repeatedly, exact once the contraction margin $m > 0$ is computed on it: weight-tied loops and deep equilibrium models are in that class, a stack of distinct layers applied once each is not.	7.1, 7.2, 7.5, 7.6 and 12.2	N31, N33 (SYN)
Attention cannot abstain unless it is given a way out	proved	exact: Unit row sums, the kernel of the softmax Jacobian (a uniform logit shift changes no weight) and the gate-times-softmax factorization hold for every head, at every layer and input.	2.7, 8.1, 8.2 and 8.4 to 8.6	N19, R7c, RIII1 (REAL)
One key among N needs a score margin growing like $\log N$	proved	exact under premises: Exact for any softmax row whose logits lie in $[-\tau, \tau]$: the bound is worst case, and the logit bound τ is read off the network.	9.1 to 9.3, 9.6, 17.2 and 17.4	R5d, V1, SX40, SX02, SX113 (REAL)
A head that cannot circulate cannot shift tokens around a cycle	proved	exact: An identity for every circulation-free head with a positive stationary law, on cycles whose successor and predecessor probabilities are positive, at any layer, stated per head; compositions of reversible heads can acquire direction, and causal heads take theirs from the mask.	10.1, 10.2, 10.5 and 10.6	N13 (REAL)
Moving information around is not enough; something must write	proved	exact: Holds for every row-stochastic mixer, at every layer and input; the Gram invariant holds for every shared orthogonal mixing of a token group.	11.1 and 11.4	ATS07, ATS12 (REAL)
We read the standard block as paying for every condition at once	interpretation	interpretation: A reading under a semi-formal price model; there is nothing to transfer.	14.1 and 14.5	AF10, AF02, BD30 (REAL)
Hardware sets the prices; the conditions are what is priced	interpretation	interpretation: A reading that joins the hardware account to the conditions; there is nothing to transfer.	14.4 and 19.6	—

Claim	Evidence	In trained networks (transfer label and note)	Sections	Records
Every machine pays to name one of N, in some currency	proved	exact under premises: Each part is exact within its class: Fano's bound and the order bound for any decoder; the logit-range (expression) bound for any soft majority; the memory floor for one-pass states with arbitrary later queries over unstructured items. The key-channel law assumes random keys on a sphere.	9.5, 19.2 and 19.3	SX29, SX57, SX58, SX59 (SYN)
Relaxing a constraint buys a capability and obliges a re-supply	interpretation	interpretation: A reading of the lattice; the individual re-supply obligations are proved in their chapters where stated.	14.2 and 14.3	—
Trained networks buy back what a component lacks	measured	measured: A mechanism measured in trained networks: necessity theorems bind exactly on their constructions, and trained networks evade them when a resource is spare.	6.6 and 15.1 to 15.3	N24b3, N12b, N12, BD30, E04b, SX48 (REAL)
No capacity of bounded gain buys back short paths	proved	exact under premises: Holds for any stack whose layer maps restrict to a common discriminative subspace with norms at most one and whose contraction product tends to zero, with the recovery gain uniformly bounded (a Dobrushin version covers sink-like attention with shared feature maps); these premises are checked layer by layer.	5.5	N25 (REAL)
Making a protection explicit does not beat a block that has it	measured	measured: Measured on twelve preregistered interventions in standard trained models.	15.4	AF01, AF02, AF04, AF08, AF10 (REAL)
Deficits may cost less as models grow	pending	pending: A hypothesis about trained networks at scale; its test is pending.	15.6	—
Under concave prices, the best design is a corner	proved	exact under premises: A theorem about allocations under concave prices; its premise, concave pricing of protections at the level of architectures, is assumed, not measured.	16.1 to 16.4	N35, E5.4 (SYN, REAL)
Read backward, adopted changes re-supply the same protections	interpretation	interpretation: A reading of adopted changes; there is nothing to transfer.	16.5	—
Read forward, the conditions rebuild a transformer	interpretation	interpretation: A reading of an assembled constraint set.	16.6	N24, N24b2 (REAL)
A 22M-parameter accuracy margin has not yet been replicated	partial	measured: Recorded at 22M parameters with 5 paired seeds against untuned baselines; not evidence of dominance until the tuned replication returns.	16.7	E2.1, SX49 (REAL)

Claim	Evidence	In trained networks (transfer label and note)	Sections	Records
No fixed rule keeps both abstention and retrieval at long context	proved	exact under premises: The impossibility bound is exact for any softmax row with content logits in $[-\tau, \tau]$ and a null shared by absent and present queries; tempering is guaranteed to keep both sides at every length only under strictly separated margins, which are checked on the head, and over a finite range it can keep both without them, as in the pilot.	17.1, 17.2 and 17.6	SX40, SX02, SX113, SX114 (REAL)
Bounded memory buys length robustness and pays in recall	proved	exact under premises: Exact for linear memories (through the rank of the keys) and for delta-rule writes with bounded values and decay under the stated step-size condition, for any sequence of keys bounded in norm.	18.1, 18.2, 18.4 and 18.6	ATS04, N36, SX47, ATS11 (REAL)
Decisions on maxima need no length schedule	partial	idealized: Shown for constructed keys and backgrounds (independent or exchangeable scores).	17.4 and 17.5	SX02 (SYN, REAL)
What makes a good writer is still open	pending	pending: Open: there is no theory yet to transfer.	11.7, 19.6 and 21.3	—
What any successor must still supply	interpretation	interpretation: A checklist read from the stated classes; there is nothing to transfer.	19.1 and 19.6	—
Most controlled evidence is small	interpretation	interpretation: A statement of scope.	21.2	—
'Cheap' rests on a price model, not measured prices	interpretation	interpretation: A statement of scope.	14.5 and 21.2	—
Explaining the past is not evidence	interpretation	interpretation: A statement of scope.	16.5 and 21.2	—
What the question uncovered (Section 0.4)				
Every unmasked attention head splits into preference, exchange and circulation†	proved	exact: The split is an identity for every Markov kernel with a positive stationary law; for every unmasked softmax head with finite scores, its score asymmetry also splits into a sink and a cycle, at every layer and input. Causal heads have no positive stationary law.	3.1 to 3.4 and 10.3	—
Trained attention's direction is specific, and it matters	measured	measured: Measured in trained encoders.	10.3 and 10.4	RCE02AAB, RCE05 (REAL)
Standard diagnostics of a single operator cannot see which way attention points	proved	exact under premises: Exact, for stationary-weighted quadratic readouts, for one amplified operator and for palindromic words of layers that share a positive stationary law. A stack of distinct layers is not a palindrome, and its readouts generally move at first order (measured median log-log slope 1.000 in the circulation strength, against 2.000, for 48 matched pairs of non-palindromic and palindromic five-layer words of balanced BERT layers from eight inputs).	3.5, 3.6 and 10.2	RC3 (REAL)

Claim	Evidence	In trained networks (transfer label and note)	Sections	Records
Depth is a distribution: short paths decide what survives	proved	idealized: Proved for path laws over powers of one fixed operator; exact for propagation on a fixed graph, not for the full nonlinear computation of a trained network, the scope its source paper states.	2.1 to 2.3, 2.5 and 2.8	—
The expected number of mixing layers sets how fast depth forgets	proved	idealized: Stated for expected energies of residual stacks with random feature maps and mean-field normalization.	2.4 and 2.7	(SYN)
Deeper trained language models send their surplus attention to the sink	partial	measured: Measured on rollout path laws, a linearization of attention flow built from each model's own attention and residual norms.	2.5 and 2.7	SX60 (REAL)
Attention rollout has an exact probabilistic meaning†	proved	exact: Exact for ordered products of causal stochastic kernels (the first token absorbing) with per-token residual shares, whatever the layers: layer-varying by construction. It describes the rollout linearization, not the full block.	8.3	(REAL)
A sharp, machine-checked stability test for adaptive averaging	proved	exact under premises: Exact for any anchored adaptive iteration once the contraction margin $m > 0$ is computed on it: weight-tied loops, deep equilibrium models, iterative denoisers and test-time mixing. A stack of distinct layers applied once each is not in the class.	7.1	—
A settled average can amplify changes and destroy accuracy	measured	measured: Measured in iterated systems that apply one builder repeatedly (softmax kernels, a real denoiser, test-time mixing over 32-image contexts); the maxima were selected from grids.	12.1 to 12.4	OUB04, OUB05 (REAL)
Each prediction's class margin splits into its own and its companions'†	proved	exact: An identity for any mixing followed by an affine readout, at every iteration and at the fixed point.	12.4 to 12.8	OUB05 (REAL)
Converging is not the same as being right or stable	partial	exact: A negative result that holds for any solver: settling certifies none of response, uniqueness, attraction or outer-loop stability; shown by explicit constructions.	13.1 to 13.10	OUB01, OUB12, OUB13 (SYN)
One stability threshold links logit scale to weight decay	proved	exact under premises: The loop-gain bound holds for every softmax layer with tied dot-product scores (projected scores multiply it by the projections' norms); the cap assumes that decoupled weight decay acts as the anchor (step share $\alpha = \eta$ and anchor share $\beta = \eta\lambda$ at learning rate η and weight decay λ), a modelling assumption.	7.3, 7.4 and 7.7	FB9, SX95 (REAL)
Softmax's $\log N$ law: abstention and retrieval cannot both be fixed	proved	exact under premises: Exact for any softmax row with content logits in $[-\tau, \tau]$ and a null shared by absent and present queries, at any layer; tempering needs strictly separated margins, checked on the head.	9.1, 9.2, 17.1, 17.2 and 17.4	SX40, SX112 (REAL)

Claim	Evidence	In trained networks (transfer label and note)	Sections	Records
Sums need a length schedule; maxima do not	partial	idealized: The argmax bound needs a uniformly placed target and a joint likelihood that depends on the target only through a nondecreasing factor of its own score; independent backgrounds with a monotone likelihood ratio satisfy this, exchangeable ones need not. The onset law assumes independent keys; the tests are constructed.	8.4, 9.3, 9.4 and 17.5	(SYN)
Soft routers pay for order out of their score range	proved	exact: A combinatorial bound for any soft router whose positional logit term has range B; its cost to the score margin, when every position must stay retrievable, is a mean-field statement.	10.7	—
Small attention leaks can grow with depth; exact block zeros stop them	proved	idealized: The growth bound is stated for path laws of one operator with a per-row crossing probability; exact block zeros stay zero for any stack whose layers share the blocks.	6.4 and 6.5	(SYN)
A fixed memory's recall is limited by capacity, however far back	proved	exact under premises: Exact for linear read-back from a d_k-dimensional memory (through the rank of the keys) and, for the state bound, for writes that contract by a common factor.	18.3	ATS04 (REAL)
Signed attention is already inside your transformer	partial	exact: A representation identity: two stochastic heads give any signed kernel with constant row sums.	15.5	BD30, BD32, BD33 (REAL)
Invertible blocks keep capability; the tested orthogonal transport loses it	partial	measured: Measured on trained mixers; the invariant behind it is exact (shared orthogonal transport preserves each token group's channel Gram matrix), but it does not imply the retrieval failure.	11.2 to 11.5	ATS07, ATS12, ATS13 (REAL)
Production vision transformers grow sinks that are not empty	measured	measured: Observational measurement in pretrained models.	8.5	SX44 (REAL)
A strong signal is not the same as a useful one	measured	measured: Measured with frozen features and label propagation on neighbour graphs.	4.4	E01e, E01d (REAL)
What this changes (Section 0.6)				
Predictions stated before the evidence	pending	pending: Predictions about trained networks, stated before their tests; the tests already run are reported with their outcomes.	20.1	—
Decoders and the MLP are only partly covered	interpretation	interpretation: A statement of scope.	10.8, 11.7 and 21.2	—
The account says nothing outside its class or its prices	interpretation	interpretation: A statement of scope.	21.2	—
The decisive tests (Section 20.4)				
Do a production model's learned sinks stop abstaining with length?	measured	measured: Prespecified on gpt-oss-20b: undecided by its prespecified rule; descriptively, its slopes lean toward compensation through position.	17.3 and 20.2	SX01 (REAL)

Claim	Evidence	In trained networks (transfer label and note)	Sections	Records
Which abstention rules keep both sides in trained language models?	measured	measured: Prespecified 125M language-model test: its lead prediction (that $\log N$ tempering on top of YaRN holds abstention and lowers degradation) failed. Every rule that gives attention a null removed the first-token sink and cut massive activations, all but sigmoid lowering perplexity at the training length; positional interpolation carried most of YaRN's gain past the training length, and log-N tempering helped on interpolated positions. In its vision subset (ViT-S/16 on ImageNet-100), tempering raised accuracy at untrained token counts without restoring the trained attention. The abstention results describe attention weights, not behaviour.	17.4 and 20.4	SX02, SX114 (REAL)
What fraction of attention layers does a qualified recurrent hybrid need?	pending	pending	18.4 and 20.4	—
Does the minimal block's accuracy margin survive tuned baselines?	pending	pending	16.7 and 20.4	—
Does compensation grow with scale?	pending	pending	15.6 and 20.4	—
Can any capacity replace a removed anchor?	pending	pending	5.5 and 20.4	—

The numbers quoted in Bruce (2026i), whose results this book prints mostly in Chapters 1–4, are listed here with their sources and tiers: **ID**, a machine-precision check of an identity (certifies algebra and code); **SYN**, a law checked on synthetic constructions; **SYN-R**, a law checked on real operators or spectra along an artificial perturbation; **REAL**, a measurement of trained models or real data. Table B.2 lists experiments of the companion record (Bruce, 2026j), each preregistered, with a falsifier where the experiment tested a prediction (RIII1 was registered as a census and E01e as a discriminator between two explanations); the verdict column gives the registered outcome, and the last column the section of this book that prints the result. Their result records, preregistrations and drivers are in the package of Bruce (2026i), in `data/` and `experiment_code/` (Appendix D). Tables B.3 and B.4 list the checks shipped with that paper.

Table B.2: Companion experiments quoted in Bruce (2026i), with the section of this book that prints each result.

ID	What	Result	Tier	Registered verdict	Where
E3.2	collapse law on graph spectra (1,500-node subgraphs of Cora, CiteSeer, PubMed, roman-empire); signed filters	Cora $\beta = 0.02612$ (lazy walk); realized filters match the closed form to 1.5×10^{-14} (collapse-ratio identity 6.5×10^{-16}); truncated restart bounded by 1.235; signed filters (synthetic, $\mu_C = 0.9$) $\kappa_{\text{col}} \approx 1$ with deep interior suppression, min $g'' = -24.2$ at $\theta = 4$	ID, SYN-R, SYN	pass, 69/75 cells; the 6 others are heterophilic cells whose predicted onset lies beyond the tested depth	Sections 2.3 and 4.1
N29	signed filters for relevant modes at low or negative eigenvalues	signed vs. nonnegative filters at matched budgets: Texas +24.3, Wisconsin +28.2, Squirrel +3.9, Actor +1.9 pp	REAL	main prediction held (the companion's cross-cell assessment, not included); rewiring-mediation sub-claim unsupported	Section 2.3
E5.3	dispersion signs	0 violations over 600,000 $(\beta, \bar{k})$ pairs	SYN	pass	Section 2.3
N30	deep-MLP control	no cross-sample collapse without a mixing operator, depth 4 only (8–64 did not converge)	REAL	pass	Section 4.1
RIII1	causal census	2,700 operators (2,544 heads and 156 head-averaged layers), 7 checkpoints; median $\delta = 0.416$; 0.67% below 10^{-6}	REAL	census (no falsifier)	Section 2.7
RIII3	sink law, 156 per-layer operators	identity 2.2×10^{-15}; envelope never violated; anchored floor held; median half-collapse 1 (interquartile range $[1, 2]$)	ID, SYN-R	pass	Section 2.7
E3.3	two readouts at small flow	270 operators (the first 24 of GPT-2 large's 36 layers), each averaged over heads and inputs, 16 checkpoints; quadratic readout ≈ 2.000 on the nine non-autoregressive checkpoints, pooled 1.905 $[1.885, 1.924]$; angle between spectral projectors pooled 1.001 $[1.001, 1.002]$; autoregressive 1.74–1.85	SYN-R	registered bands (1.94–2.0; 0.96–0.98) missed; 'pass' only under the driver's $[1.6, 2.4]$, $[0.6, 1.4]$	Section 3.6
RIII2	causal fingerprint	identity 4.9×10^{-16} (156 layers); $\sum m_{\text{off}} / \sum m_{\text{self}} = 7.89$ (mean of per-layer ratios 9.82), $\Phi_{\text{F}} = 0.942$ at the pooled ratio; encoders fail (median identity residual 0.114, 54 layers)	ID, REAL	mixed	Section 3.6
E3.1	family fingerprints	causal family 0.860 $[0.842, 0.879]$ (per layer and head; π-reversibilized parts, plain Frobenius norms)	REAL	recorded 'pass' under a scoping added after the freeze; under the frozen preregistration (a) and (b) failed (13 and 3 violations, all causal)	Section 3.6
E3	two-mode contamination	contamination multiplied by R (slope 1.000000 in $\log R$)	ID	pass	Section 4.3

ID	What	Result	Tier	Registered verdict	Where
N6	damage against the smoothing cap	slope 1.002 in R; up to 495× the cap	SYN	pass	Section 4.3
E1.2	multiclass fragility identity	1.5×10^{-15}; post-order damage zero	ID	pass	Section 4.3
E01	no task band	median-bandwidth Gaussian graphs, frozen CIFAR features, 3 backbones × 2 datasets × 10 seeds (60 cells), 40 labels per class; task gap 4.7×10^{-4} (N09 gauge; record not included); at $\alpha = 0.9$ both orders lose in every cell, 21–61 pp (medians −41 post, −40 pre; means per backbone and dataset 34–57)	REAL	inconclusive as recorded (60/60); one-sidedness not reproduced, both orders harm	Section 4.4
E01e	label density	CIFAR-100 k-NN graphs, 15; per graph, the smallest change among post-placement runs: −0.1 to −3.9 pp (80% labels) vs. +7.6 to +10.9 pp (8%); paired +10.9 pp $[10.2, 11.6]$ at 8% vs. 80% (registered primary, both low densities: +8.8 pp $[8.2, 9.4]$); monotone in density on 15/15; the gate admitted the 15 high-density cells and none of the 30 low-density cells	REAL	discriminator; returned the headroom branch	Section 4.4
E01d	deflated gate on real kernels; in-harness controls	real features (3 backbones × 2 datasets × 5 seeds): deflated $R > 1$ in 24/30 sharp-kernel cells, 2/30 (k-NN, $k = 10$), 0/30 ($k = 20$); gate rejects all sharp-kernel cells; in the 24, at $\alpha = 0.5$, pre −0.8 to −4.3 pp, post −4.3 to −25.7 pp, pre never worse; 16/26 gate-passing cells with decisive post harm; positive control (nuisance $\mu = 0.99$ above class modes at 0.9): post +1.2 to +7.4 pp at $\alpha = 0.3$–0.7, pre −1.8 to −6.4 pp at $\alpha = 0.9$, decisive pre harm in 5/5	REAL, SYN	falsifier fired (real cells); controls passed	Sections 4.3 and 4.4
E1.1	harm search	66 cells; R defined on 24, eight distinct spectra (the other 42 are subsampled graphs left reducible or with a degenerate stationary distribution); max $R = 0.9965$ (Texas, $N = 183$, at $\alpha = 0.1$; margin $\mu_n - \mu_d = -0.032$); feature kernels 0.97–0.99; identity 1.9×10^{-15}	REAL, ID	registered null held	Section 4.3
E5.2	misalignment bound	0/2,000 violations, one synthetic substrate per check; worst ratio 0.075; mean slopes 1.997 vs. 0.862 (median 0.961)	SYN	pass	Section 4.3

Table B.3: Captured-operator checks shipped with Bruce (2026i) (`code/real_operators.py`).

ID	What	Result	Tier
RC2	GPT-2 small, causal operators	2,304 per-head operators: median $\delta = 0.080$ (interquartile range $[0.020, 0.216]$), 2.2% below 10^{-6}; 192 per-layer operators: median $\delta = 0.29$, identity 6.7×10^{-16}, 0 envelope and floor violations, half-collapse median 2 (max 8); median composite distance 4.7×10^{-5}, 0.090, 0.76, 1.28 at $b = 1, \frac{1}{2}, 0.2, 0.1$, 0 violations of Proposition 2.20(i); fingerprint identity 6.7×10^{-16}, Φ_F mean 0.877 (median 0.954), 0.948 at the mean per-layer ratio 8.82	ID, REAL, SYN-R
RC3	BERT-base and DeiT-small at native flow	192 / 96 lazy per-layer operators; Φ median 0.061 / 0.059 (max 0.14); odd part $\leq 1.3 \times 10^{-15}$; collapse-type change median 0.38% / 0.29% (restart), 0.27% / 0.17% (residual stream), 1.5% / 1.1% (depth 4; max 5.9%); bound over change, median 30 / 32 (restart), 4.5 / 5.4 (residual stream), 5.2 / 6.0 (depth); sin median 0.29 / 0.28 (max 0.92 / 0.96); measured over first-order prediction, median 1.07 / 1.06 (within the window: 1.02 on the one BERT operator, median 1.01 on the 18 DeiT operators); window $\Phi < \gamma_C/2$: 1 of 192 / 18 of 96; local gap median 0.032 / 0.049; $d_f(P)/d_f(S) < 1$ for restart and residual stream (max 0.9999993); depth median 1.0012 / 1.0013 (range 0.9994–1.0054 / 1.00002–1.011); $\pi_{\max}/\pi_{\min}$ median 391 / 378; effective support 5.8% / 12.9% of tokens; small-t slopes 2.000 and 1.00	ID, REAL
RC4	words of balanced BERT layers	48 palindromes: slope 2.000 (range 1.99995–2.0001), odd part at native flow $< 5.1 \times 10^{-16}$; 48 non-palindromes: slope 1.000 (0.974–1.015), odd part median 7.6%; balanced layers' Φ median 0.131	SYN-R

Table B.4: Checks of the verification suite of Bruce (2026i) (`code/verify_theory.py`, RC1) that its text cites, with their numbers.

Check	Statement	Numbers	Tier
X1	two-community chain, $\beta = 0.05$	depth 2.27, 5.16, 26.6 ($L = 16, 32, 64$); restart 1.45; stream 1.221; share $\frac{1}{2}$: 5.05	ID
X2	lopsided triangle	eigenvalues $\{1, 0.3366, 0.1634\}$; P_s row-stochastic for $\lvert s\rvert \leq 0.2$; $\Phi = \sqrt{3}\,\lvert s\rvert$	ID
X3	self-or-sink, $\delta = 0.1$	0.957, 0.197 ($L = 7, 22$); half-collapse 6.6; stack 1.69 ($L = 10$) $\to$ 1.637; floor 0.271	ID
S3.4	Theorem 2.8: products, window, species prices	depth $\leq$ stack $\leq$ stream $\leq$ restart at matched suppression on 36 grid points; closed forms to 7×10^{-15}	ID
S3.7	boundary of Theorem 2.8	$\kappa_{\text{col}} = 1$; suppression 100; price 1.144; five negative coefficients	ID
S3.9	Theorem 2.16(iv) scope	$\lVert Q\rVert_1 = 1.8$	ID
S3.10	Theorem 2.16(iii) scope	short-path mass without an anchor: distance 0 at a fixed operator	ID
S3.12	Theorem 2.18, Proposition 2.19	0 violations over 2,884 operators, six families (conditioning variant at $m = 1, 2$)	SYN
S3.13	anchor floor in between	two-token swap with $a = \frac{1}{2}$: $f(P) = \mathbf{1}\pi^\top$ exactly	ID
S3.14	leak and direction	leak of P_s: $0.65 - s$; of P_{-s} at $s = 0.1$: 0.60; ℓ_∞ distances of $f(P)$ and $f(P^\dagger)$ differ in 142 of 200 random cases	ID, SYN

Check	Statement	Numbers	Tier
S3.16	Definition 2.17 remarks	$\lambda = \delta$ on causal operators; directed witness $\lambda = 0.780 > \bar{\gamma} = 0.772$	ID
S3.17	Poisson tail	$\Pr[K \geq m] \leq e^{-\sigma}(e\sigma/m)^m$ for $m \geq \sigma$; fails below ($\sigma = 4$, $m = 1$: 0.982 against 0.199)	SYN
S3.18	Theorem 2.8 on families	anchor-free $\{1, \theta\}$ lottery: $\kappa_{\mathrm{col}} \leq 2.11$	SYN
S3.19	Proposition 2.9	one-mode price against a linear program: error 10^{-13}; proof inequality of (ii): 0 violations on 2,563 (polynomial, m) pairs from 233 random polynomials; binomial witnesses meet both bands	SYN
S4.11	Proposition 3.7(iii)	coefficient error 1.6×10^{-9}; $\{0, 10\}$ lottery: $+8.91$ untilted, -7.16 tilted	ID
S4.12	Proposition A.3	0.3352 and 0.3418	ID
S4.13	Proposition 3.7(iv)	0 violations over 7,500; depth ratio as low as 0.67	SYN
S4.14	Proposition A.4	collision at $\Phi = 0.0866$, half the gap	ID
S4.15	Lemma 3.8	$\Phi_{\mathrm{F}}^2 = 1/7$ vs. $1/9$; $\Phi = 0.173$ while Φ_{F}: $0.229 \to 0.044$	ID
S4.16, S4.17	Proposition 3.7(i)	non-lazy $d_f = 0.2 + \Phi$; lazy slopes 2.000 (444 cases and a degenerate one)	ID, SYN
S4.19	Table 3.2	reversal-invariant statistics even to 6.1×10^{-14}; small-flow slopes 2.000 (smooth) and 1.000 (norms of odd objects, cross readouts)	SYN
S4.20	degenerate plane (Section 3.5)	eigenvalues $\lambda \pm ict$ to 10^{-16}; characteristic polynomial even	ID
S4.21	registration rule (Bruce, 2026i, Appendix E)	secant bound: 0 violations on 2,000 random draws (those with $\eta \geq 1$ skipped; 1,605 tested); E3.3 window: $\eta \leq 0.125$ for ± 0.06	SYN, ID
S5.2	Theorem 4.1	identity 2.1×10^{-15}; no random map with $Ey = 0$ exceeds $\varepsilon\|r\|_\pi$ (0/1,200); directed escape derivative matches $\langle y, [g(S), \mathcal{D}_g]y\rangle_\pi$ to 10^{-11}, zero for a normal operator	ID, SYN
S5.3	Corollary 4.2	0 violations over 300 lazy reversible operators; turn bound attained (ratio 1) on the lazy reversible walk on a four-token ring; depth 200 turns a seed with $w = 0.01$ by 84.3°	SYN, ID
S5.4	Theorem 4.3	commutator 5.6×10^{-17}, deviation 0.024	ID
S5.6	Theorem 4.1(iv)	$\langle n, Y_{\mathrm{pre}}\rangle_\pi = 0$; $\langle n, Y_{\mathrm{post}}\rangle_\pi = 0.150$ (depth 1), 0.030 (depth 4)	ID

For the decisive tests still pending, to which Chapter 0 refers here, the register is Table 20.1 in Section 20.4: what each test measures, its status, and what each of its outcomes would change in the account of Chapter 0. It is generated from the same claims export as Table B.1, whose last rows name the same tests, and it is not repeated here.

How the collection registered its experiments and judged their outcomes, and how it checked its statements, stay in the source papers' own appendices on evidence and validity, for example Bruce (2026d, Appendix B) and Bruce (2026i, Appendix E). Section R.2 defines the evidence and transfer labels that the claims ledger carries, and the tiers are defined above, with the evidence ledger of Bruce (2026i).

Appendix C

Notation

Table C.1 gathers the book's notation: every symbol that recurs across its chapters, with its meaning and the chapters that use it. It extends the notation of Bruce (2026i) to the whole book, in the names that Section R.4 fixes.

Table C.1: **Notation.** Every symbol that recurs across the chapters, with its meaning and the chapters that use it in that meaning; where a row gives several symbols, each listed chapter uses at least one of them. R is the reading guide, and a chapter's entry includes its proofs in Appendix A. A letter with several meanings has one row for each, and a chapter may also give a letter a meaning of its own, defined where it is used and not listed here. Chapter 0 prints its source unedited; the rows say where its symbols differ. Standard notation ($\mathbb{E}$, Pr or $\mathbb{P}$, I, $\mathbf{1}$, e_i, diag, $\|\cdot\|_2$, $\|\cdot\|_F$, $^\top$, softmax) is not listed.

Symbol	Meaning	Chapters
Latin letters		
A	the anchors: fixed input representations that an anchored iteration re-injects at every step	0, 1, 4, 7, 12, 13
A	the reciprocal traffic $\frac{1}{2}(F + F^\top) = D_\pi S$, the symmetric part of the stationary flow	R, 3, 10
A_ℓ	the attention kernel of layer ℓ in a stack (head-averaged in a rollout)	1, 2, 8
a	the anchor $\Pr[K = 0] = f(0)$: the share of the output that is the unpropagated input (p_0 in Chapter 6)	1, 2, 4–6, 21
a, w	a head's abstention, its weight on the null for an absent query, and its weight on the match for a present one; a_N, w_N at N keys	8, 9, 17
a	the score asymmetry $s - s^\top$	3, 10
a	the decay factor that a recurrent write applies to the state: a bounded state needs $a < 1$	0, 16, 18
a_*, w_*	the abstention and the single-key retrieval weight that a head is to keep (Chapter 0 writes a^*, w^*)	0, 9, 17
B	the background weight $\sum_j e^{s_j}$ against which a null or a match competes ($\sum_j e^{\tau s_j}$ at logit scale τ); B_N at N keys	8, 9
B, r, K	the range $\max_j b_j - \min_j b_j$ of a positional term in the logits; the ratio $r > 1$ by which one write is to outweigh another; $K = \lfloor B/\ln r \rfloor$	0, 9, 10
B_J	the update that a layer's circulation delivers, $\sum_h J_h V_h W_O^{(h)}$ over its heads h	3, 10
B_ℓ	in a causal rollout, layer ℓ's factor on the tokens after the first; its products give the survival $\|B_L \cdots B_1 \mathbf{1}\|_\infty$	8, 21
b, z	the null (sink) logit, and the null's weight $z = e^b$ (z_N, z_i); under a length schedule $b = b_0 + \ln N$	7–9, 17

Symbol	Meaning	Chapters
b_j, $b(\lvert i-j\rvert)$	a bias added to a logit: a key bias b_j, which is a potential; a positional term; a relative-position bias $b(\lvert i-j\rvert)$	3, 10
b_ℓ	the update share of layer ℓ in $(1-b_\ell)I + b_\ell P_\ell$, and b for a constant share (see p_ℓ)	1, 2, 4, 5
C	the current $F - F^\top = 2D_\pi J$, twice the antisymmetric part of the stationary flow	R, 3, 10
C	the number of classes	0, 6, 15, 16
c	the score of the target (matching) key; in the constructions, its cosine with the query	6, 8–10, 17
D, W	a nonnegative affinity matrix W and the diagonal matrix D of its row sums (degrees), with $P = D^{-1}W$	1, 2, 6, 19
D	the dimension of constructed random keys	6, 9, 10
$D(R\Vert Q)$	the stationary divergence rate $\sum_i \mu_i D_{\mathrm{KL}}(R_{i\cdot}\Vert Q_{i\cdot})$ of kernels R and Q, with μ the stationary law of R	3, 10
$\mathcal{D}_f$	the first-order array of $f(P_t)$ at $t=0$, with modal entries $f[\mu_k,\mu_l]J_{kl}$ ($\mathcal{D}_g$ for an amplifier g)	3, 4
D_π	$\mathrm{diag}(\pi)$	R, 1–4, 10
DF	the Jacobian of an update: of F_A at an equilibrium X^*, or DF_t along a trajectory; the reprinted figures of Chapters 12 and 13 write it J	R, 7, 12, 13
d	the width, the feature dimension of $X \in \mathbb{R}^{N\times d}$	1, 2, 5, 7, 12, 13, 18
d_b	the dimension of the content that keys carry	0, 6, 15
d_k	the key dimension of a head; a head's linear memory stores at most d_k associations exactly	0, 1, 18, 21
d_{TV}	the total-variation distance	1, 2
E, Q	for two tokens, the projections $E = \mathbf{1}\mathbf{1}^\top/2$ onto their common value and $Q = I - E$ onto their contrast	7, 12, 13
F	the stationary flow $D_\pi P$: F_{ij} is the probability that a stationary step goes from token i to token j	3, 10
F	the distribution function of a background score	6, 8, 9
F_A	the anchored update $X \mapsto \alpha\mathcal{P}(X)X + \beta A$, with iterates $X_{t+1} = F_A(X_t)$	1, 7, 12, 13
G, q_ℓ	the gain (Lipschitz constant) of a map that recovers what a constrained state retains; $G \geq c/\prod_\ell q_\ell$ when the layers of a stack act on a common subspace with norms $q_\ell \leq 1$, and $A_\epsilon = \mathrm{diag}(1,\epsilon)$ is injective but needs gain $1/\epsilon$	0, 5, 15
G	a rank-one gain, which uniformly random unit probes read as $G/\sqrt{n}$ in root mean square	0, 13
g, f	an amplifier $g(z) = \sum_k c_k z^k$ with $c_k \geq 0$, and its normalized form $f = g/g(1) = \mathbb{E}[z^K]$, so that $f(P) = \mathbb{E}[P^K]$	0–6, 10, 14
g	a gate, the factor by which a head scales a softmax row, $P = \mathrm{diag}(g)\,\mathrm{softmax}_{\mathrm{row}}(s)$; $g = 0$ abstains	8, 9, 14
g	the inner initializer of a finite computation, $X_0 = g(A)$	7, 13
g	in the two-token constructions, the builder's contrast function: $\mathcal{P}(X) = E + g\big((x_1 - x_2)/2\big)Q$	7, 12, 13
H	the symmetrically normalized kernel $D^{-1/2}WD^{-1/2}$	2, 6
H	the centering matrix $I - \mathbf{1}\mathbf{1}^\top/N$	3, 5
H	the response $D_A X^* = \beta(I - DF)^{-1}$ of an equilibrium to its anchors	7, 12, 13
H_K, K	the derivative of K inner updates of a finite computation (its finite derivative)	7, 13

Symbol	Meaning	Chapters
H_π	the stationary conditional entropy	3, 10
h	the cross-community weight of the two-community chain, a test walk of Section 1.5	1, 2
$\mathcal{I}$	the circulation information $\sum_i \pi_i D_{\mathrm{KL}}(P_{i\cdot}\|S_{i\cdot}) = D(P\|S)$, in nats	3, 10
J	the circulation $\frac{1}{2}(P - P^\dagger)$, π-antisymmetric: the flow, or routing, of Chapters 1, 3 and 6; the reprinted figures of Chapters 12 and 13 write J for an update's Jacobian DF	R, 1, 3, 6, 10, 19, 21
K, p_k	the path length, a random number of steps (the lottery, or path law), and its law $p_k = \Pr[K = k]$ (T in Chapter 6)	0–3, 5, 6, 8
K	the number of loops of a weight-tied block: K_{train} in training, $K_{1/2}$ where accuracy has halved	5, 13
K0, K2	the tied dot-product softmax builder, and the Gaussian softmax builder with adaptive bandwidth	7, 12, 13
k	the levels $1/k$ of a hybrid's attention fraction, one attention layer in k ($1/K$ in Chapter 20)	14, 16, 20
L, ℓ	the depth (the number of layers), and the layer index	0–6, 8, 14–16, 20, 21
L_P	the Lipschitz constant of the builder on B_R	0, 1, 7, 12, 13
M	a second linear map on the token index, a readout or a projection	1, 4, 6
M	the number of samples that a sampling router (a walker) draws	8, 9, 19
M, N	a weight-tied loop of M distinct blocks, each visited N times	2, 20, 21
m	the certified contraction margin $\beta - \alpha\kappa_{\mathrm{f}}$; β/m bounds the response of the equilibrium to its anchors	0, 1, 7, 12, 13
m	the drift of a residual branch's map, $V_\ell = m_\ell I + Z_\ell$	2, 21
m	a short-path window: the path law keeps mass at least ρ on lengths at most m	2, 5
$m(\tau)$	the mean weight $\mathbb{E}e^{\tau z}$ of a background score z at logit scale τ	6, 8–10
N	the number of tokens; for one query, the number of keys it competes over, the context length. Published rules named after it write n (the $-\log n$ bias of sigmoid attention, $\log n$ scaling), and a causal sequence has T tokens in Chapters 2 and 8	0–10, 12–21
N_*	the exact simultaneous budget $e^{2\tau}(1 - a_*)(1 - w_*)/w_*$: the most background keys at which a null that absent and present queries share keeps both targets (Chapter 0 writes N^*)	0, 9, 17
N^*	the length at which a weight or an accuracy first fails, its onset; in the mean field, the majority onset $e^{\tau c}/m(\tau)$	9, 19
N_{train}, N_0	the number of keys that a head was trained at	0, 9, 17, 21
P	a row-stochastic operator on the tokens: an attention head's kernel $P(X)$, rebuilt from the representations X; a graph's propagation matrix; a kernel smoother	0–8, 10–15, 18, 19
$P^\dagger$	the time reversal of P, its π-adjoint, with entries $\pi_j P_{ji}/\pi_i$	R, 1–3, 8, 10
P_t, P_α	the sweep $S + tJ$ from $P^\dagger$ ($t = -1$) through S ($t = 0$) to P ($t = 1$); Chapters 3 and 10 write the dose α for t	1, 3, 10
P_s, s	the lopsided triangle $P_s = S + sC$, a test walk of Section 1.5, with rotation strength s and C the circulation on three tokens	1–3
P_*	the weights $\mathcal{P}(X^*)$ realized at an equilibrium	4, 12
$\mathcal{P}$	the builder, which returns nonnegative weights $\mathcal{P}(X)$ whose rows sum to one (Chapter 0 writes $P(X)$)	0, 1, 4, 7, 12, 13
p_ℓ, p	the share of the averaging (attention) branch in layer ℓ's update, $p_{\ell,i}$ for token i: the update share b_ℓ	2, 5, 8, 14, 20, 21

Symbol	Meaning	Chapters
Q	a substochastic matrix: the transient block of a strictly causal walk, or the weights of a head that can abstain	2, 7, 8
Q, K	query and key feature matrices, with scores $s = \lambda QK^\top$	3, 10
q	a contraction factor: of an anchored iteration, $\alpha(1+\kappa_{\mathrm{f}}) = 1-m$, or $\alpha(r_{\max} + RL_Q)$ for weights whose row mass is at most $r_{\max}$; in Chapter 18, of a recurrent write	7, 12–14, 18
R, B_R	a radius that bounds the representations in the maximum-row norm, and the ball $B_R = \{X : \|X\|_{\infty,2} \le R\}$	R, 0, 1, 7, 12, 13, 16, 19, 20
R	the resolvent smoother $\beta(I-\alpha P)^{-1} = (1-\alpha)(I-\alpha P)^{-1}$ of a fixed P, the restart's lottery operator	R, 5, 12
R	the relative energy that a mode keeps through a stack, $R_L(\mu)$ or $R_{M,N}(\mu)$	R, 2
R	the dynamic range $g(\mu_n)/g(\mu_d)$ between a nuisance mode and a readout mode	4, 21
R	the width of a soft read's logit window; in Chapter 19, the ratio of its largest to its smallest weight	9, 19
R^2	the coefficient of determination of a retrieval or a fit	0, 6, 10, 11, 15
R_*	the frozen smoother $\beta(I-\alpha P_*)^{-1}$	4, 7, 12
r	the row mass $\sum_j p_j$ of a head that can abstain, $1-r$ its abstention; $r_{\max}$ a bound on it	7, 8, 14
r	the cycle term HaH of the scores	3, 10
r	a restart's re-injection probability, $r = 1-\alpha$	2, 20, 21
S	the reversible part $\frac{1}{2}(P + P^\dagger)$, who averages with whom: preference and exchange	R, 1–4, 10, 19, 21
S_L	the mixing budget $\sum_\ell \Pr[K_\ell \ge 1]$, the expected number of layers that mix, with K_ℓ the steps that layer ℓ takes	0, 2, 5
s, s_{ij}	attention scores (logits) before the softmax, s_j for one query	1, 3, 6–10, 17
s, s_c	the logit scale of a dot-product builder, as in $\kappa_{\mathrm{f}} \le 2sR^2$ (see τ); $s_c = \lambda/(2R^2)$, its critical value when weight decay acts as the anchor	0, 7, 12, 13, 16, 20
$s_{\ell,i}$	the attention that token i pays to the first token, the sink, at layer ℓ	2, 8
t	a diffusion time, in the heat filter $e^{t(P-I)}$	2, 6, 21
t	a species' strength: the depth L, the total update σ, or $\alpha/(1-\alpha)$ for a restart	14, 16
U	uniform averaging, $\mathbf{1}\mathbf{1}^\top/N$	6, 12, 14
v_a, v_r, u_r	the margins $N_0 e^{s_0-b}$, $N_0 e^{s_0-s_*}$ and e^{b-s_*} of the tempering theorem, for background scores at most s_0 and a match score s_*	9, 17
W_Q, W_K, W_V, W_O	a head's query, key, value and output maps, and W_{QK} its query–key score form	0, 1, 3, 6, 10, 11, 14, 15
W_{sk}	the skew part $\frac{1}{2}(W - W^\top)$ of a bilinear score matrix W	3, 10
w_k, $\mathrm{sd}_w(g)$	a seed's spectral weights, and the gain spread, the standard deviation of the gain under w	1, 4
X^*	an equilibrium (fixed point) of an anchored iteration	7, 12, 13
Y_{pre}, Y_{post}	the outputs of the two orders for a seed Y: $Y_{\mathrm{post}} = g(P)MY$ (project, then amplify) and $Y_{\mathrm{pre}} = Mg(P)Y$ (amplify, then project)	R, 1, 4, 6
y_i, $y_h(i)$	the sink's log-odds against the background at query position i, $b - \mathrm{LSE}_j(s_{ij})$, for head h in Chapter 17	9, 17

Symbol	Meaning	Chapters
Greek letters		
α	a restart's continuation probability (the geometric lottery); the weight on the averaged term of an anchored iteration $X_{t+1} = \alpha\mathcal{P}(X_t)X_t + \beta A$	0–7, 12–14, 16, 18–20
α_ℓ	the scale of layer ℓ's residual branch relative to the stream (α_L when it depends on the depth)	2, 21
β	$1-\alpha$, the anchor's share: the weight on the anchor or restart	0, 1, 4, 7, 12–14, 18–20
β	the content gap $1-\mu_C$, the fraction of readout content that one averaging step destroys	1, 2, 4, 5, 14, 16
γ, $\bar\gamma$	the spectral gap $1-\mu_2$ ($\bar\gamma$ for the reversible part of a directed P)	1–3
γ	a retention level: a mode μ is kept at gain $f(\mu) \geq \gamma$	2, 21
γ	the strength of an outer feedback loop around a finite computation	7, 13
γ	the growth of a head's background log-sum per unit of $\log n$; a stationary background has $\gamma = 1$	9, 17, 20
δ	the sink weight $\min_{i\geq 2} P_{i1}$ of a strictly causal operator	1, 2, 4, 8
ϵ	the one-step leak of a filter across a block boundary	0, 6
ε	the step of a weight-tied block's update at each visit	2, 21
ζ	an eigenvalue of P, complex in general ($\zeta = \xi + i\eta$ in Chapter 3)	2, 3
η	the share that a single key may lose to its competitors: it keeps $1-\eta$	0, 6, 9
η; μ_-, μ_+	a nuisance cap, $f(\mu_-) \leq \eta$ on a mode μ_-; under it, the largest gain on a signal mode $\mu_+ > \mu_-$ is $w\mu_+^k + (1-w)\mu_+^{k+1}$	2, 21
η	a learning rate; in Chapter 18, the step of a recurrent write	7, 18, 20
θ	a threshold on a head's largest score, below which it abstains	8, 9
κ_{col}	the collapse ratio $g(1)/g(\mu_C) = 1/\mathbb{E}[(1-\beta)^K]$	R, 1–5, 14
κ_{circ}	the circulation index $2\|\widetilde{J}\|_F^2/\|\widetilde{P}\|_F^2$	R, 3, 10
κ_{f}	the feedback constant RL_P, at most $2sR^2$ for dot-product softmax (Chapter 0 writes κ_f)	R, 0, 1, 7, 12, 14, 16
$\lambda(P)$	the uniform leak, the smallest overlap of two rows of P; $1-\lambda(P)$ is Dobrushin's coefficient, which Chapter 5 writes $\tau(P)$	2, 5
λ	the scale of the scores in $s = \lambda QK^\top$	3, 10
λ	the symmetric fraction of a weight-tied block's map (λ_{s} in Chapter 20)	2, 20, 21
λ	a weight decay; in Chapter 18, the decay of a recurrent state	0, 7, 18, 20
λ_N	the tempering factor $\log N/\log N_0$ that multiplies every score, the null included (λ in Chapter 9)	9, 17
μ, μ_k, v_k	an eigenvalue (a mode) of S, the argument of a gain $f(\mu)$; the eigenvalues $1 = \mu_1 \geq \cdots \geq \mu_N$, with π-orthonormal eigenvectors v_k and $v_1 = \mathbf{1}$	R, 0–6, 12, 14, 20, 21
μ	the mean of the background exponentials e^{s_j}, in the mean-field law $a_{\mathrm{mf}} = z_N/(z_N + N\mu)$	R, 9
μ_C	the readout (content) eigenvalue: the largest eigenvalue among the content modes v_k, $k \geq 2$, that a readout depends on	2, 3, 5, 14, 16, 20
μ_0, Σ	an interior mode $0 < \mu_0 < \mu_C$ that a design suppresses, with gap $\beta_0 = 1-\mu_0$ (β_n in Chapter 16), and its suppression $\Sigma = g(1)/g(\mu_0)$	2, 14, 16
π	the stationary law, $\pi^\top P = \pi^\top$, positive and unique when P is irreducible	1–4, 6, 8, 10, 12, 14, 19, 21
ρ	the mass that a path law keeps on lengths at most a window m	2, 5

Symbol	Meaning	Chapters
ρ	the spectral radius $\rho(\cdot)$; for a substochastic head Q, its Perron root, with $\|Q\|_w = \rho < 1$	2, 8, 10, 12–14
σ	the total update of a residual stream, the mean of its Poisson path length	1–4, 14, 16
σ	the noise scale of a residual branch's map, the standard deviation of its centered part Z_ℓ	2, 21
σ	the entropy production $D_{\mathrm{KL}}(F\|F^\top)$ of the stationary flow	3, 10
τ	a bound on the content logits, which lie in $[-\tau, \tau]$	R, 0, 9, 17, 20
τ	the logit scale, a gain that multiplies the scores (weights $e^{\tau s_j}$, logits τc_j); learned as $\tau_{\max}\sigma(\theta)$ in Chapter 17, with $\tau_{\max}$ a bound on it. For cosine scores it also bounds the logits; compare s	R, 0, 5–10, 17, 19
τ	a temperature that divides the logits, $q^\top k_j/\tau$	6, 15
τ_i	the time that the walk from token i takes to reach the absorbing token 1	2, 8
Φ	the flow ratio $\|J\|_\pi \in [0, 1]$	1, 3
Other notation		
$\langle u, v\rangle_\pi$, $\|\cdot\|_\pi$	the π-weighted inner product $\sum_i \pi_i \bar{u}_i v_i$ and its norm, with the induced operator norm on matrices; $L^2(\pi)$ is the space they define	1–4, 6, 8, 14
$\langle u, v\rangle$	the plain (Euclidean or Frobenius) inner product	1–3, 6, 7, 10, 12, 13
$\|\cdot\|_\infty$	the largest absolute row sum (the ℓ_∞ operator norm), at most 2 for a difference of row-stochastic matrices	1, 2, 7, 8, 12, 15, 21
$\|\cdot\|_{\infty,2}$	the maximum-row norm $\max_i \|x_i\|_2$, written $\|\cdot\|_\infty$ in the proofs for Chapter 5	1, 5, 7, 12
$M^\dagger$	the π-adjoint $D_\pi^{-1} M^\top D_\pi$ of a matrix M; for P, the time reversal $P^\dagger$	1, 3, 6
$\widetilde{M}$	stationary coordinates $D_\pi^{1/2} M D_\pi^{-1/2}$, in which $\widetilde{S}$ is symmetric and $\widetilde{J}$ skew	3, 10
$\mathbf{1}e_1^\top$	full collapse onto the first token: the matrix whose every row is $e_1^\top$	1, 2, 8
log, ln	the natural logarithm; both are written	0–3, 5, 6, 8–10, 14–21
LSE	log-sum-exp, $\mathrm{LSE}(s) = \log \sum_j e^{s_j}$, and LSE_i for row i	8, 9, 17
σ, sigmoid	the logistic function, written sigmoid in Chapter 15	0, 8, 9, 15, 17

Each chapter defines its local symbols where they first appear, and Table C.1 lists only those that recur; Section R.4, which points here, states the choices and the clashes that a reader meets first.

Appendix D

Records and packages

Each source paper's package holds the records and code behind its results, and Table D.1 lists the packages: for each paper, its folder in the collection, its vendored record sets, its experiment code, the builders of its figures and tables, and its Lean proofs where it has them. The packages follow this text in a later version of the collection, each archived as its own Zenodo record and cross-linked to this book. The book's own numbers and tables are generated by the `paper_code/` of its own package: `make_book_numbers.py` computes the simple aggregates of records that the chapters print and no paper states (seed means and ratios, with their differences, ranges and counts) and the values they read directly from a record, checking each record file against its manifest first, and `make_book_tables.py` writes the book's versions of the papers' tables, among them Tables B.1 and 20.1, and Table D.1 itself, read from the packages. The ten figures of Bruce (2026h) that Chapters 12 and 13 reprint are copied byte for byte into `book/figures/p02/`, whose manifest lists each file's SHA-256.

Table D.1: **The source papers' packages.** For each paper the book draws on: its folder in the collection, the number of its vendored record sets (Sets), its experiment code, the builders of its figures and tables, and its Lean proofs, read from the package itself. In every package each record file is listed with its SHA-256 in a checksum manifest (`MANIFEST.sha256`), and the package's `README.md` describes its contents.

Paper	Package	Sets	Code	Builders	Lean
The Transformer Abides (Chapter 0)	`P0.0_why_transformers/paper/`	8	`experiment_code/`	`paper_code/`	—
The Arrow of Attention (Bruce, 2026g)	`P0.1_reciprocity_and_circulation/paper/`	6	`experiments/`	`make_all.sh`	`lean/`
When Averaging Amplifies (Bruce, 2026h)	`P0.2_when_averaging_amplifies/paper/`	9	`experiment_code/`	`paper_code/`	`lean/`
A Theory of Spectral Amplification (Bruce, 2026i)	`P0.4_theory_of_spectral_amplification/paper/`	17	`experiment_code/`	`paper_code/`	—
Depth Is a Distribution (Bruce, 2026k)	`P0.5_depth_is_a_distribution/paper/`	6	`experiment_code/`	`paper_code/`	—
Attention Sinks at Long Context (Bruce, 2026c)	`P06_softmax_logN_laws/paper/`	46	`experiment_code/`	`paper_code/`	—
The Transformer as a Constrained Equilibrium (Bruce, 2026d)	`CAP_transformer_constrained_equilibrium/paper/`	64	`experiment_code/`	`paper_code/`	—
Beyond Transformers (Bruce, 2026e)	`LAT_lattice_machines/paper/`	15	`experiment_code/`	`paper_code/`	—
Signed Attention Is Already Inside Your Transformer	`P02_signed_attention/paper/`	23	`experiment_code/`	`paper_code/`	—
Invertible Is Not Isometric	`P03_invertible_not_isometric/paper/`	16	`experiment_code/`	`paper_code/`	—

Paper	Package	Sets	Code	Builders	Lean
Signal Is Not Utility	`P04_signal_not_utility/paper/`	14	`experiment_code/`	`paper_code/`	—
Mechanisms Are Not Knobs	`P05_mechanisms_not_knobs/paper/`	17	`experiment_code/`	`paper_code/`	—
Own and Companion Margins	`P08_own_and_companion_margins/paper/`	4	`experiment_code/`	`paper_code/`	—
Non-Expansive Routing	`P09_nonexpansive_routing/paper/`	9	`experiment_code/`	`paper_code/`	—
The Compensation Principle	`P10_compensation_principle/paper/`	9	`experiment_code/`	`paper_code/`	—
Coupling Logit Scale to Weight Decay Removes a Class of Self-Referential Instabilities	`P11_logit_scale_weight_decay/paper/`	4	`experiment_code/`	`paper_code/`	—
Settled Is Not Solved	`P12_settled_not_solved/paper/`	5	`experiment_code/`	`paper_code/`	—
A Minimal Anchored Block vs Parameter-Matched ViTs	`P13_minimal_anchored_block/paper/`	5	`experiment_code/`	`paper_code/`	—
Recurrent Memory vs Attention at Matched Data	`P14_recurrent_memory_vs_attention/paper/`	6	`experiment_code/`	`paper_code/`	—
Why Architectures Sit at Corners	`P16_architectures_at_corners/paper/`	6	`experiment_code/`	`paper_code/`	—
The Leak Cliff	`P17_leak_cliff_exact_sparsity/paper/`	1	`experiment_code/`	`paper_code/`	—
Certifying Self-Referential Feedback Loops	`P18_coupled_feedback_certification/paper/`	6	`experiment_code/`	`paper_code/`	—

Table D.2 gives the design record with its sources: every system in the two records of deployed designs that Section 16.3 reads, the curated table and the later record (Bruce, 2026d), with the values each records, the primary source of each value, a released configuration file or a technical report, the dates on which that source was read, and how each review flag of the curated table was resolved. The flagged values of MiniMax-01 and Ouro are confirmed, that of Jamba-1.5 only for its predecessor, and those of Gated DeltaNet-H1 and Huginn-3.5B remain unconfirmed; the later record leaves those two out. The attention fraction of Kimi Linear carried no flag, but the curated table rounds it, and the later record keeps the value of the released configuration.

Table D.2: **The design record with its sources.** Every system in the two records of deployed designs that Section 16.3 reads: the curated table (Curated), whose 13 rows carry review flags on five, and the later record of six systems (Later), each read from a released configuration file or a technical report. Checked gives the dates on which each primary source was read for the value; the curated table records no date of its own. The last column gives each review flag and how it was resolved. A dash marks a system that a record does not list.

System	Axis	Curated	Later	Primary source	Checked	Review flag and resolution
Qwen3-Next-80B-A3B	Attention fraction	1/4	12/48	Released configuration file and model card (Qwen Team, 2025)	June 2026; 30 Sep. 2026	None.
Kimi Linear 48B-A3B	Attention fraction	1/4	7/27	Released configuration file; technical report (Kimi Team, 2025)	June 2026; 30 Sep. 2026	None. The curated table rounds the nominal ratio of the technical report; the later record keeps the value of the released configuration.
Gated DeltaNet-H1	Attention fraction	1/2	—	Technical report (Yang et al., 2025)	Value not checked	The interleave of this variant, to be confirmed. Unresolved; not in the later record.

System	Axis	Curated	Later	Primary source	Checked	Review flag and resolution
MiniMax-01	Attention fraction	1/8	—	Technical report and released configuration file (MiniMax, 2025)	30 Sep. 2026	Outside the source paper's bibliography. Confirmed; not in the later record.
Jamba-1.5	Attention fraction	1/8	—	Technical report of its predecessor, Jamba (Lieber et al., 2024)	30 Sep. 2026	Outside the source paper's bibliography. Confirmed for Jamba, not checked for Jamba-1.5; not in the later record.
Ouro-1.4B	Loop count	4	4	Technical report (Zhu et al., 2025)	June 2026; 8 Oct. 2026	The trained loop count, to be confirmed. Confirmed; the later record keeps one row for the Ouro models.
Ouro-2.6B	Loop count	4	4	Technical report (Zhu et al., 2025)	June 2026; 8 Oct. 2026	None. The later record keeps one row for the Ouro models.
Huginn-3.5B	Loop count	32	—	Technical report (Geiping et al., 2025)	Value not checked	Trained with a randomized depth, so the value is a mean. Unresolved; not in the later record.
ViT-Tiny	Depth	12	—	The ViT lineage, no specific source	Not checked	None. Not in the later record.
ViT-Small	Depth	12	—	The ViT lineage, no specific source	Not checked	None. Not in the later record.
ViT-Base	Depth	12	12	Technical report, Table 1 (Dosovitskiy et al., 2021)	June 2026	None.
ViT-Large	Depth	24	24	Technical report, Table 1 (Dosovitskiy et al., 2021)	June 2026	None.
ViT-Huge	Depth	32	32	Technical report, Table 1 (Dosovitskiy et al., 2021)	June 2026	None.

The protocols of the experiments, and how a package's numbers, figures and tables are rebuilt from its records, stay with the source papers: in their own appendices on protocols and reproducibility, such as the appendix on reproducibility of Bruce (2026g) and that on experimental details and records of Bruce (2026c), and in each package's `README.md`.

Bibliography

Laziz U. Abdullaev and Tan M. Nguyen. Transformer meets twicing: Harnessing unattended residual information. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.00687.

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4895–4901, 2023.

Habibullah Akbar. ATMA: Long-context language modeling via polar attention and gated-delta compression memory. arXiv preprint arXiv:2606.25156, 2026.

David Aldous and James Allen Fill. Reversible Markov chains and random walks on graphs. Unfinished monograph, recompiled 2014, 2002. URL https://www.stat.berkeley.edu/~aldous/RWG/book.html.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations (ICLR)*, 2021.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society and Oxford University Press, Providence, RI, 2000. doi: 10.1090/mmono/191. URL https://doi.org/10.1090/mmono/191.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*, 2024a.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1763–1840. PMLR, 2024b.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3:133–181, 1922. doi: 10.4064/fm-3-1-133-181.

Arpit Bansal, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. End-to-end algorithm synthesis with recurrent networks: Extrapolation without overthinking. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20232–20242. Curran Associates, Inc., 2022.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael Bronstein. Why do LLMs attend to the first token? In *Conference on Language Modeling*, 2025a.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://arxiv.org/abs/2410.06205.

Alexander Bendikov and Laurent Saloff-Coste. Random walks on groups and discrete subordination. *Mathematische Nachrichten*, 285(5–6):580–605, 2012. doi: 10.1002/mana.201000059.

Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, 2000.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, volume 202, pp. 2397–2430, 2023.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Advances in Neural Information Processing Systems*, 2023.

Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *International Conference on Learning Representations (ICLR)*, 2024.

Viacheslav Borovitskiy, Iskander Azangulov, Alexander Terenin, Peter Mostowsky, Marc Deisenroth, and Nicolas Durrande. Matérn gaussian processes on graphs. In *International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2593–2601, 2021. URL https://proceedings.mlr.press/v130/borovitskiy21a.html.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2105.14491.

Neil D. B. Bruce. A spectral theory of token communication: Operators, amplification and the architecture of transformers. Zenodo, version 0.1, 2026a. URL https://doi.org/10.5281/zenodo.23541723.

Neil D. B. Bruce. Attention sinks at long context: One log-N law for retrieval and abstention. Zenodo, released with this book, 2026c. URL https://doi.org/10.5281/zenodo.23541963.

Neil D. B. Bruce. The transformer as a constrained equilibrium: A lattice of protections for adaptive sequence mixing, and where the room is. Zenodo, released with this book, 2026d. URL https://doi.org/10.5281/zenodo.23542051.

Neil D. B. Bruce. Beyond transformers: Sums, maxima and the price of order. Zenodo, released with this book, 2026e. URL https://doi.org/10.5281/zenodo.23542133.

Neil D. B. Bruce. Lost in ranking: When attribution scores hide class contrast. Manuscript, 2026f.

Neil D. B. Bruce. The arrow of attention: Sinks, cycles, and circulation in transformers. Manuscript, 2026g.

Neil D. B. Bruce. When averaging amplifies: Stability and sensitivity in adaptive mixing. Manuscript, 2026h.

Neil D. B. Bruce. A theory of spectral amplification: Collapse, direction, and the placement of gain. Manuscript, 2026i.

Neil D. B. Bruce. The experimental program of "A theory of spectral amplification": Protocols, findings, and implications. Technical report, 2026j.

Neil D. B. Bruce. Depth is a distribution: Path laws and the mixing budget of deep propagation. Manuscript, 2026k.

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Larry Carter, Robert Floyd, John Gill, George Markowsky, and Mark Wegman. Exact and approximate membership testers. In *Proceedings of the Tenth Annual ACM Symposium on Theory of Computing*, pp. 59–65, 1978.

Ben Chamberlain, James Rowbottom, Maria I. Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. GRAND: Graph neural diffusion. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1407–1418, 2021. URL https://proceedings.mlr.press/v139/chamberlain21a.html.

Bo Chang, Minmin Chen, Eldad Haber, and Ed H. Chi. AntisymmetricRNN: A dynamical system view on recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Fang Chen, László Lovász, and Igor Pak. Lifting Markov chains to speed up mixing. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pp. 275–281, 1999.

Jie Chen, Weiqi Liu, Zhizhong Huang, Junbin Gao, Junping Zhang, and Jian Pu. Universal Deep GNNs: Rethinking Residual Connection in GNNs from a Path Decomposition Perspective for Preventing the Over-smoothing, 2022. URL https://arxiv.org/abs/2205.15127. arXiv:2205.15127.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *International Conference on Machine Learning (ICML)*, 2020a.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. arXiv:2306.15595, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020b. URL https://proceedings.mlr.press/v119/chen20j.html.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7654–7664. Association for Computational Linguistics, 2022.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized PageRank graph neural network. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=n6jl7fLxrP.

Aakanksha Chowdhery et al. PaLM: Scaling language modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828/.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations (ICLR)*, 2020.

Martin Courtois, Malte Ostendorff, Leonhard Hennig, and Georg Rehm. Symmetric dot-product attention for efficient training of BERT language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8002–8011, 2024.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454. URL https://doi.org/10.1214/aop/1176996454.

Róbert Csordás, Christopher D. Manning, and Christopher Potts. Do language models use their depth efficiently? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Wentao Cui and Yuhong Guo. Parameterless transductive feature re-representation for few-shot learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2212–2221, 2021. URL https://proceedings.mlr.press/v139/cui21a.html.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359. Curran Associates, Inc., 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers need registers. In *International Conference on Learning Representations*, 2024.

Soham De and Samuel L. Smith. Batch normalization biases residual blocks towards the identity function in deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434, 2024.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345): 118–121, 1974. doi: 10.1080/01621459.1974.10480137.

Mostafa Dehghani et al. Scaling vision Transformers to 22 billion parameters. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512. PMLR, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: CompleteP enables compute-efficient deep transformers. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.01618.

Persi Diaconis, Susan Holmes, and Radford M. Neal. Analysis of a nonreversible Markov chain sampler. *The Annals of Applied Probability*, 10(3):726–752, 2000.

Roland L. Dobrushin. Central limit theorem for nonstationary Markov chains. I, II. *Theory of Probability and Its Applications*, 1(1, 4):65–80, 329–383, 1956.

Wolfgang Doeblin. Sur les propriétés asymptotiques de mouvements régis par certains types de chaînes simples. *Bulletin Mathématique de la Société Roumaine des Sciences*, 39:57–115 (no. 1); 3–61 (no. 2), 1937. Published in two parts.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. FlexAttention: A programming model for generating fused attention variants. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 7, 2025. URL https://proceedings.mlsys.org/paper_files/paper/2025/hash/61a9278dfef5f871b5e472389f8d6fa1-Abstract-Conference.html.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Nick Dungey. Subordinated discrete semigroups of operators. *Transactions of the American Mathematical Society*, 363(4):1721–1741, 2011.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019.

Nelson Elhage et al. A mathematical framework for Transformer circuits. Transformer Circuits Thread, https://transformer-circuits.pub/2021/framework/index.html, December 2021.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin, 1997.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 12606–12633, 2024.

Katie Everett, Lechao Xiao, Mitchell Wortsman, Alexander A. Alemi, Roman Novak, Peter J. Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling exponents across parameterizations and optimizers. In *International Conference on Machine Learning (ICML)*, 2024.

William Feller. *An Introduction to Probability Theory and Its Applications*, volume I. Wiley, New York, 3rd edition, 1968.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Alexander M. Fichtl, Jeremias Bohn, Josefin Kelber, Edoardo Mosca, and Georg Groh. The end of Transformers? On challenging attention and the rise of sub-quadratic architectures. arXiv preprint arXiv:2510.05364, 2025. Revised version published as "Challenging Quadratic Attention – A Holistic View On the Rise of Alternative Language Model Architectures", Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 60–81, Association for Computational Linguistics, 2026.

James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *The Annals of Applied Probability*, 1(1):62–87, 1991.

G. David Forney, Jr. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Transactions on Information Theory*, 14(2):206–220, 1968.

Fabrizio Frasca, Emanuele Rossi, Davide Eynard, Ben Chamberlain, Michael Bronstein, and Federico Monti. SIGN: Scalable inception graph neural networks, 2020. URL https://arxiv.org/abs/2004.11198. arXiv:2004.11198.

Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3–4):193–206, 1990. doi: 10.1080/0022250X.1990.9990069.

Guoji Fu, Peilin Zhao, and Yatao Bian. p-Laplacian based graph neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6878–6917, 2022. URL https://proceedings.mlr.press/v162/fu22e.html.

Peng Gao et al. Lumina-T2X: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. arXiv preprint arXiv:2405.05945, 2024.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=H1gL-2A9Ym.

Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 13333–13345, 2019b.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41340–41391. Curran Associates, Inc., 2025.

Gemma Team. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021.

Francesco Di Giovanni, James Rowbottom, Benjamin P. Chamberlain, Thomas Markovich, and Michael M. Bronstein. Understanding convolution on graphs via energies. *Transactions on Machine Learning Research*, 2023.

Siddharth Gollapudi, Nilesh Gupta, Prasann Singhal, and Sewon Min. Can language models actually retrieve in-context? Drowning in documents at million token scale. arXiv preprint arXiv:2607.01538, 2026.

Winfried K. Grassmann, Michael I. Taksar, and Daniel P. Heyman. Regenerative analysis and steady state distributions for Markov chains. *Operations Research*, 33(5):1107–1116, 1985.

Alessio Gravina, Davide Bacciu, and Claudio Gallicchio. Anti-symmetric DGN: a stable architecture for deep graph networks. In *International Conference on Learning Representations (ICLR)*, 2023.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3748–3758, 2020. URL https://proceedings.mlr.press/v119/grazzi20a.html.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg K. H. Franke, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *International Conference on Learning Representations*, 2025.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The unreasonable ineffectiveness of the deeper layers. In *International Conference on Learning Representations (ICLR)*, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Fangda Gu, Heng Chang, Wenwu Zhu, Somayeh Sojoudi, and Laurent El Ghaoui. Implicit graph neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2009.06211.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, 2025.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs. *arXiv preprint arXiv:2410.13835*, 2024. URL https://arxiv.org/abs/2410.13835.

Dongchen Han, Tianyu Li, Ziyi Wang, and Gao Huang. Vision transformers are circulant attention learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(26):21549–21557, 2026. doi: 10.1609/aaai.v40i26.39303. URL https://ojs.aaai.org/index.php/AAAI/article/view/39303.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *Conference on Language Modeling (COLM)*, 2024.

Soufiane Hayou, Eugenio Clerico, Bobby He, George Deligiannidis, Arnaud Doucet, and Judith Rousseau. Stable ResNet. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1324–1332, 2021. URL https://proceedings.mlr.press/v130/hayou21a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Mingguo He, Zhewei Wei, Zengfeng Huang, and Hongteng Xu. BernNet: Learning arbitrary graph spectral filters via Bernstein approximation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648, 2020. doi: 10.1145/3397271.3401063. URL https://arxiv.org/abs/2002.02126.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253. Association for Computational Linguistics, 2020.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision (ECCV)*, pp. 289–305, 2024.

Jordan Hoffmann et al. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022.

Sara Hooker. The hardware lottery. *Communications of the ACM*, 64(12):58–65, 2021.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin R. Benson. Combining label propagation and simple models out-performs graph neural networks. In *International Conference on Learning Representations*, 2021.

Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating Gaussian diffusions. *The Annals of Applied Probability*, 3(3):897–913, 1993.

Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating diffusions. *The Annals of Applied Probability*, 15(2):1433–1444, 2005. doi: 10.1214/105051605000000025. URL https://doi.org/10.1214/105051605000000025.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21502–21521. PMLR, 2024.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127:203–244, 2011.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Nikita Karagodin, Shu Ge, Yury Polyanskiy, and Philippe Rigollet. Normalization in attention dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive Transformers with linear attention. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020.

Tosio Kato. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Springer, 1995.

Ali Kayyam, Anusha Madan Gopal, and M. Anthony Lewis. Do transformers need three projections? systematic study of QKV variants. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2606.04032.

Frank P. Kelly. *Reversibility and Stochastic Networks*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Chichester, 1979. URL https://www.statslab.cam.ac.uk/~frank/BOOKS/kelly_book.html.

John G. Kemeny and J. Laurie Snell. *Finite Markov Chains*. Springer, 1976.

Jack Kendall, Ross Pantone, Kalpana Manickavasagam, Yoshua Bengio, and Benjamin Scellier. Training end-to-end analog neural networks with Equilibrium Propagation. arXiv preprint arXiv:2006.01981, 2020.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *International Conference on Machine Learning (ICML)*, 2021.

Kimi Team. Kimi Linear: An expressive, efficient attention architecture. arXiv preprint arXiv:2510.26692, 2025. Model card: https://huggingface.co/moonshotai/Kimi-Linear-48B-A3B-Instruct.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient Transformer. In *International Conference on Learning Representations*, 2020.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating residual and normalization layers into analysis of masked language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

A. Kolmogoroff. Zur Theorie der Markoffschen Ketten. *Mathematische Annalen*, 112(1):155–160, 1936. doi: 10.1007/BF01565412. URL https://doi.org/10.1007/BF01565412.

Md Kowsher, Nusrat Jahan Prottasha, Chun-Nam Yu, Ozlem Garibay, and Niloofar Yousefi. Does self-attention need separate weights in transformers? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 535–543, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-industry.44. URL https://aclanthology.org/2025.naacl-industry.44/.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Maxime Labonne. Qwen3.5: Nobody agrees on attention anymore. Hugging Face community article, https://huggingface.co/blog/mlabonne/qwen35, February 2026.

Lucien Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197, 1960.

Vincent Leroy, Jerome Revaud, Thomas Lucas, and Philippe Weinzaepfel. Win-win: Training high-resolution vision transformers from two windows. In *International Conference on Learning Representations*, 2024.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, 2nd edition, 2017. With contributions by Elizabeth L. Wilmer and a chapter by James G. Propp and David B. Wilson.

Shuzhen Li, Yifan Zhang, Jiacheng Guo, Quanquan Gu, and Mengdi Wang. DeepLoop: Depth scaling for looped transformers. *arXiv preprint arXiv:2607.13491*, 2026. URL https://arxiv.org/abs/2607.13491v2. Version 2.

Opher Lieber et al. Jamba: A hybrid Transformer-Mamba language model. arXiv preprint arXiv:2403.19887, 2024.

Siyi Lin, Chongming Gao, Jiawei Chen, Sheng Zhou, Binbin Hu, Yan Feng, Chun Chen, and Can Wang. How do recommendation models amplify popularity bias? an analysis from the spectral perspective. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2025.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020a.

Meng Liu, Hongyang Gao, and Shuiwang Ji. Towards deeper graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 338–348, 2020b. doi: 10.1145/3394486.3403076.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986. URL https://doi.org/10.1109/ICCV48922.2021.00986.

Enzhe Lu et al. MoBA: Mixture of block attention for long-context LLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17790–17815, 2025.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Physical Review Letters*, 132(7):077301, 2024. doi: 10.1103/PhysRevLett.132.077301.

Pierre Marion, Adeline Fermanian, Gérard Biau, and Jean-Philippe Vert. Scaling ResNets in the large-depth regime. *Journal of Machine Learning Research*, 26(56):1–48, 2025. URL https://jmlr.org/papers/v26/22-0664.html.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Sohir Maskey, Raffaele Paolino, Aras Bacho, and Gitta Kutyniok. A fractional graph laplacian approach to oversmoothing. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.13084.

William J. McCausland. Time reversibility of stationary regular finite-state Markov chains. *Journal of Econometrics*, 136(1):303–318, 2007.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 20192–20204, 2025. doi: 10.18653/v1/2025.findings-acl.1035.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

William Merrill and Ashish Sabharwal. The expressive power of Transformers with chain of thought. In *International Conference on Learning Representations*, 2024.

Milena Mihail. Conductance and convergence of Markov chains: A combinatorial treatment of expanders. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 526–531, 1989.

Evan Miller. Attention is off by one. Blog post, https://www.evanmiller.org/attention-is-off-by-one.html, July 2023.

MiniMax. MiniMax-01: Scaling foundation models with lightning attention. arXiv preprint arXiv:2501.08313, 2025.

L. Mirsky. A dual of Dilworth's decomposition theorem. *The American Mathematical Monthly*, 78 (8):876–877, 1971. doi: 10.2307/2316481.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. Quantifying context mixing in transformers. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2023.

Trishit Mondal and Ameya D. Jagtap. In transformer we trust? A perspective on transformer architecture failure modes. arXiv preprint arXiv:2602.14318, 2026.

Ken M. Nakanishi. Scalable-softmax is superior for attention. arXiv preprint arXiv:2501.19399, 2025.

Sharan Narang et al. Do Transformer modifications transfer across implementations and applications? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5758–5773. Association for Computational Linguistics, 2021.

Tam Nguyen, Tan M. Nguyen, and Richard G. Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/48b383b24230e0e6e649d9c98dae4d8c-Abstract-Conference.html.

J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.

Andrew O'Desky. On attention heads and bilinear forms. *arXiv preprint arXiv:2609.22990*, 2026. URL https://arxiv.org/abs/2609.22990.

Catherine Olsson et al. In-context learning and induction heads. Transformer Circuits Thread, https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html, March 2022. Also arXiv:2209.11895.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations (ICLR)*, 2020.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. Model card, https://openai.com/index/gpt-oss-model-card/, August 2025. Released 5 August 2025; also arXiv:2508.10925.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 18724–18741, 2024.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66, Stanford InfoLab, November 1999. URL http://ilpubs.stanford.edu:8090/422/. Previous number SIDL-WP-1999-0120.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations*, 2024.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *International Conference on Learning Representations (ICLR)*, 2023.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

Zihan Qiu et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, volume 38, pp. 100092–100118. Curran Associates, Inc., 2025.

Qwen Team. Qwen3-Next: Towards ultimate training & inference efficiency. Blog post, https://qwen.ai/blog?id=4074cca80393150c248e508aa62983f9cb7d27cd, September 2025. Model card: https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Jason Ramapuram, Federico Danieli, Eeshan Gunesh Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russell Webb. Theory, analysis, and best practices for sigmoid self-attention. In *International Conference on Learning Representations*, 2025.

Edward T. Reehorst and Philip Schniter. Regularization by denoising: Clarifications and new interpretations. *IEEE Transactions on Computational Imaging*, 5(1):52–67, 2019. doi: 10.1109/TCI.2018.2880326. URL https://phil-schniter.web.app/pdf/tci19_red.pdf.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017. doi: 10.1137/16M1102884. URL https://elad.cs.technion.ac.il/wp-content/uploads/2018/02/LittleEngine-Final.pdf.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.07524.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3515–3530. PMLR, 2022.

Matteo Saponati, Pascal Josef Sager, Pau Vilimelis Aceituno, Thilo Stadelmann, and Benjamin F. Grewe. The underlying structures of self-attention: Symmetry, directionality, and emergent dynamics in transformer training. In *International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 52958–52994, 2025.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear Transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366. PMLR, 2021.

J. Schnakenberg. Network theory of microscopic and macroscopic behavior of master equation systems. *Reviews of Modern Physics*, 48(4):571–585, 1976. doi: 10.1103/RevModPhys.48.571.

Damien Scieur, Gauthier Gidel, Quentin Bertrand, and Fabian Pedregosa. The curse of unrolling: Rate of differentiating through optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17133–17145, 2022. doi: 10.52202/068431-1246. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6d53193a098b982229340a7c3eb0ecbf-Abstract-Conference.html.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. The MultiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.16163.

Eugene Seneta. *Non-negative Matrices and Markov Chains*. Springer Series in Statistics. Springer-Verlag, New York, 2nd edition, 1981.

Claude E. Shannon. Probability of error for optimal codes in a Gaussian channel. *Bell System Technical Journal*, 38(3):611–656, 1959.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. DeltaProduct: Improving state-tracking in linear RNNs via Householder products. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964.

Alexander J. Smola and Risi Kondor. Kernels and regularization on graphs. In *Learning Theory and Kernel Machines*, pp. 144–158, 2003. URL https://alex.smola.org/papers/2003/SmoKon03.pdf.

Jianlin Su. Viewing the scale operation of attention from the perspective of entropy invariance. Scientific Spaces (blog), https://kexue.fm/archives/8823, December 2021. In Chinese; English translation of the original title.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling*, 2024.

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to Transformer for large language models. arXiv preprint arXiv:2307.08621, 2023.

Zhong Sun, Giacomo Pedretti, Elia Ambrosi, Alessandro Bricalli, Wei Wang, and Daniele Ielmini. Solving matrix equations in one step with cross-point resistive arrays. *Proceedings of the National Academy of Sciences*, 116(10):4123–4128, 2019. doi: 10.1073/pnas.1815682116.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, pp. 3319–3328, 2017.

Rich Sutton. The bitter lesson. Blog post, http://www.incompleteideas.net/IncIdeas/BitterLesson.html, March 2019.

Shawn Tan, Songlin Yang, Aaron Courville, Rameswar Panda, and Yikang Shen. Scaling stick-breaking attention: An efficient implementation and in-depth study. In *International Conference on Learning Representations*, 2025.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12342–12364. Association for Computational Linguistics, 2023.

Matthew Thorpe, Tan Nguyen, Hedi Xia, Thomas Strohmer, Andrea L. Bertozzi, Stanley J. Osher, and Bao Wang. GRAND++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EMxu-dzvJk.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations (ICLR)*, 2022.

Behrouz Touri and Angelia Nedić. On ergodicity, infinite flow, and consensus in random models. *IEEE Transactions on Automatic Control*, 56(7):1593–1605, 2011. doi: 10.1109/TAC.2010.2091174. URL https://arxiv.org/abs/1001.1890.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 2021.

Lloyd N. Trefethen and Mark Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.

Paul Turán. On an extremal problem in graph theory. *Matematikai és Fizikai Lapok*, 48:436–452, 1941. In Hungarian.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Andreas Veit, Michael J. Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, 2016.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, 2025.

Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=O476oWmiNNp.

Peihao Wang, Ruisi Cai, Yuehao Wang, Jiajun Zhu, Pragya Srivastava, Zhangyang Wang, and Pan Li. Understanding and mitigating bottlenecks of state space models through the lens of recency and over-smoothing. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2501.00658.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep Transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1810–1822. Association for Computational Linguistics, 2019.

Richard Zhe Wang. Abstention and noise filtering: Two missing primitives of softmax attention. arXiv preprint arXiv:2609.22005, 2026.

Shaowen Wang, Bingrui Li, Ge Zhang, Wenhao Huang, Shen Yan, and Jian Li. On the residual scaling of looped transformers: Stability and transferability. *arXiv preprint arXiv:2606.18524*, 2026. URL https://arxiv.org/abs/2606.18524v2. Version 2.

Yifei Wang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Dissecting the diffusion process in linear graph convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

ZiDong Wang, Zeyu Lu, Di Huang, Cai Zhou, Wanli Ouyang, and Lei Bai. FiTv2: Scalable and improved flexible vision transformer for diffusion model. arXiv preprint arXiv:2410.13925, 2024.

Geoffrey Wolfer and Shun Watanabe. Information geometry of reversible Markov chains. *Information Geometry*, 4(2):393–433, 2021.

Felix Wu, Tianyi Zhang, Amauri H. de Souza Jr., Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6861–6871, 2019. URL https://proceedings.mlr.press/v97/wu19e.html.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67756–67781, 2025. URL https://proceedings.mlr.press/v267/wu25ad.html.

Louis-Pascal A. C. Xhonneux, Meng Qu, and Jian Tang. Continuous graph neural networks. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10432–10441, 2020. URL https://proceedings.mlr.press/v119/xhonneux20a.html.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Zhenda Xie et al. mHC: Manifold-constrained hyper-connections. arXiv preprint arXiv:2512.24880, 2025.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the Transformer architecture. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533. PMLR, 2020.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5453–5462. PMLR, 2018.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs VI: Feature learning in infinite-depth neural networks. In *International Conference on Learning Representations*, 2024a. URL https://arxiv.org/abs/2310.02244.

Liu Yang, Kangwook Lee, Robert Nowak, and Dimitris Papailiopoulos. Looped transformers are better at learning learning algorithms. In *International Conference on Learning Representations*, 2024b.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear Transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, volume 37, pp. 115491–115522. Curran Associates, Inc., 2024c.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025.

Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Revisiting Semi-Supervised Learning with Graph Embeddings. In *International Conference on Machine Learning (ICML)*, volume 48, pp. 40–48, 2016.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential Transformer. In *International Conference on Learning Representations*, 2025.

Jingyang Yuan et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23078–23097. Association for Computational Linguistics, 2025.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing Transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803. PMLR, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models, 2022. arXiv:2205.01068.

Wentao Zhang, Mingyu Yang, Zeang Sheng, Yang Li, Wen Ouyang, Yangyu Tao, Zhi Yang, and Bin Cui. Node dependent local smoothing for scalable graph learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2110.14377.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, 2023.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, 2003.

Hao Zhu and Piotr Koniusz. Simple spectral graph convolution. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CYO5T-YjWZV.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7793–7804, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/58ae23d878a47004366189884c2f8440-Abstract.html.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, Lu Li, Jiajun Shi, Kaijing Ma, Shanda Li, Taylor Kergan, Andrew Smith, Xingwei Qu, Mude Hui, Bohong Wu, Qiyang Min, Hongzhi Huang, Xun Zhou, Wei Ye, Jiaheng

Liu, Jian Yang, Yunfeng Shi, Chenghua Lin, Enduo Zhao, Tianle Cai, Ge Zhang, Wenhao Huang, Yoshua Bengio, and Jason Eshraghian. Scaling latent reasoning via looped language models. arXiv preprint arXiv:2510.25741, 2025.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning (ICML)*, 2003.