The Transformer as a Constrained Equilibrium: A Lattice of Protections for Adaptive Sequence Mixing, and Where the Room Is

Neil D. B. Bruce *brucen@uoguelph.ca*
University of Guelph

Abstract

A sequence-mixing layer that is iterated through depth and rebuilt from its own input faces six failure modes. Each is met by a protection—anchoring, alignment, gain control, abstention, a routing budget, or an arrow—with an operator-level statement proved within a declared class, a signature when it is removed, and a price. We organize attention designs as a lattice of kernel, dynamical, and block constraints in which relaxing a constraint buys a capability and obliges a re-supply. We read softmax attention inside a pre-normalized residual block as the lattice point where every protection is either supplied or cheaply re-supplied under parallel-hardware and optimization pricing: of the deficits tested, all but the anchor's were compensated by capacity elsewhere in the deeper models, and explicit protection knobs did not beat the block, both measured at small scale, while corner structure, by which designs would jump rather than drift, remains a hypothesis. The lattice reads the selected adopted changes to the mixing layer, its normalization and its residual path over the last decade as cheap re-supplies of protections, and it locates the room for change. The clearest room is abstention, governed by a two-sided log-N law: with a bounded logit scale, no null keeps both abstention and single-key retrieval as context grows, and tempering the entire logit budget with $\log N$ keeps both when its margins are strictly separated. A controlled pilot follows the law, while a prespecified test of its sink prediction on an open production model was undecided by its prespecified rule; descriptively its slopes leaned toward compensation through position, in which position encoding holds the background's growth down. In a prespecified trained harness of 125M-parameter 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 and raised long-context perplexity. In these language models every rule that gives attention a null removed the first-token sink and cut massive activations by 67% to 82%, as earlier reports for gates and attention biases found, and, sigmoid aside, lowered perplexity slightly at the training length; the comparison registered for two of these rules held in every seed, while the same comparison failed in the harness's vision transformers. We close with falsifiable predictions for future architectures, most of them not yet tested.

1 Introduction

Nine years after its introduction (Vaswani et al., 2017), the transformer is still the default sequence model. Many alternatives have beaten it on particular axes: linear attention and state-space models on inference cost, recurrent memories on the memory cost of long contexts, sparse attention on long-context compute. None has replaced it. The modifications that practice did adopt are small and specific: moving layer normalization to the start of each residual branch, normalizing queries and keys, adding sink tokens, registers or learned sink logits, gating attention outputs, scaling logits with the logarithm of context length, and interleaving a few attention layers with recurrent layers at fixed fractions. Why is the design so stable, and where, if anywhere, can it move?

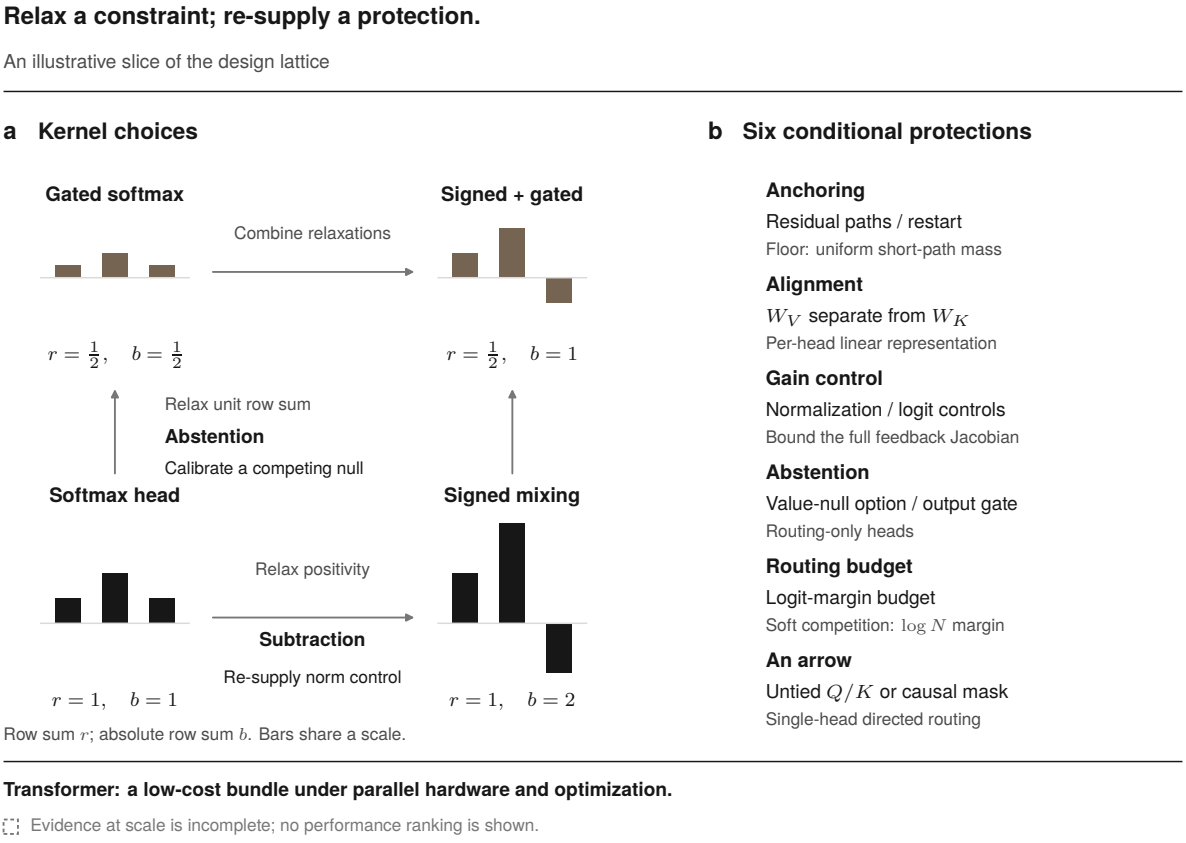


Figure 1: Attention designs are organized by the protections they preserve and the capabilities they gain. (a) An illustrative construction shows the exact Boolean lattice obtained by relaxing nonnegative weights and unit row sum. The bottom rows are $(1/4, 1/2, 1/4)$ and $(1/2, 1, -1/2)$; the upper 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 the paper's architectural interpretation under the stated price model, not a measured optimum. No empirical measurements are drawn; evidence at scale remains incomplete.

We answer by treating a mixing layer as what it is: an adaptive Markov operator applied to token representations inside an anchored residual block. A softmax head is a row-stochastic kernel $P(X)$ rebuilt from its input at every layer. Iterating and rebuilding such operators exposes six failure modes: collapse toward a consensus mode; contamination of content by routing; runaway self-reinforcement of adaptive weights; being forced to deliver an update when nothing is relevant; dilution as the number of competitors grows; and the inability to represent direction. Each failure has a protection that is proved within a declared class of models and has a signature measured when it is removed (synthetic, for gain control; Section 3). The protections are rows of a design inventory, not six mechanisms every model must implement. Within normalized nonnegative path laws, retention needs short-path mass, of which an identity path is one form. The arrow is a statement about a single head, and compositions of reversible heads can acquire direction. Abstention is a requirement for heads that can only route. Gain control concerns the full feedback Jacobian, not normalization alone.

The protections organize the design space into a lattice (Section 5). Each constraint that softmax attention satisfies—unit row sums, positivity, density, a pre-normalized residual—guarantees something and forbids something. Relaxing it buys a capability, and obliges the designer to re-supply what it guaranteed. We read softmax attention in a pre-normalized residual block as the lattice point where every protection is either supplied or cheaply re-supplied given parallel hardware and gradient-based optimization; this is an interpretation under a price model, not a measured optimum (Section 10). Two measured mechanisms help hold it there (Section 7): compensation, by which trained networks evade the deficits tested through capacity elsewhere, at least in the deeper models, the anchor being the exception (nothing made up for its loss, and under strict mixing nothing of bounded gain can); and the failure of explicit, certified protection knobs to beat a block that already pays for the protections. A third, corner structure, by which optimal designs would jump between discrete allocations rather than drift, is a hypothesis: it holds under concave prices, which are assumed, not measured.

The lattice makes predictions as well as describing designs (Section 9). The selected changes that practice adopted in the mixing layer and its block over the last decade (Table 5) read as cheap re-supplies of protections the lattice names, a reading made after the fact and so not evidence for the lattice; changes to the MLP, to the implementation and to the block's topology lie outside its scope (Section 7). It also locates where the room is (Section 8). The clearest case is abstention. Rows that sum to one over values a head cannot zero leave it no way to abstain, and every fix proposed or in use—sink tokens, off-by-one normalization, learned sink logits, sigmoid attention with a $-\log N$ bias, output gates—is a point in one cell of the lattice. A two-sided law governs the cell (Bruce, 2026c). With logits bounded by τ, no null keeps both abstention and single-key retrieval beyond an explicit N_* proportional to $e^{2\tau}$, whether it is fixed, learned, or shifted by $\log N$. Tempering the whole logit budget, sink included, with $\log N$ keeps both when the sink and the match hold strictly separated margins. In a controlled pilot, trained fixed-sink heads follow the law's mean-field prediction of abstention to within 0.002 per seed. A prespecified test of its sink prediction in a production model, gpt-oss, was undecided by its prespecified rule; descriptively its slopes leaned toward compensation through position, in which position encoding holds the background's growth down. In a prespecified trained harness of 125M-parameter 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 and raised long-context perplexity, while every rule that gives attention a null removed the first-token sink and most massive activations, as earlier reports for gates and attention biases found (Bondarenko et al., 2023; Sun et al., 2024; Qiu et al., 2025) (Section 8).

Contributions.

1. A conditional inventory of six protections for adaptive sequence mixing, each with an operator-level statement proved within a declared class, a removal signature measured in controlled experiments (synthetic, for gain control), and a price (Section 3), and the structure–content separation theorem that grounds the alignment row (Section 4).

2. A lattice of kernel, dynamical and block constraints that places attention variants, recurrent layers and hybrids in one map, with the capability each relaxation buys and the re-supply it obliges (Section 5).

3. An account of the transformer's stability as a constrained equilibrium: compensation and protection knobs that do not help, both measured, and corners, a hypothesis; plus a retrodiction of recent architectural practice (Section 7).
4. The clearest room: the two-sided log-N law of abstention and retrieval (Bruce, 2026c), with a controlled pilot, a production-model test, and a prespecified test in trained 125M-parameter language models that refuted the fix applied on top of YaRN and found, as registered for two null rules, that a null removes the first-token sink and most massive activations in these language models; the registered vision-transformer comparison failed (Section 8).
5. Falsifiable predictions for future architectures; apart from the sink length law, tested on a production model (undecided) and in the trained 125M harness (where fixed sinks took the registered compensated branch), sigmoid retrieval, tested in the same harness, and a provisional comparison for transport-only mixers, they have not yet been tested (Section 9).

What we do not claim. We do not claim a model that beats tuned transformers on standard benchmarks at scale; the fixed-point analysis predicts that successors will win mainly on resource axes. The largest controlled test is at 125M parameters, and production-scale evidence consists of measurements on public checkpoints, among them a prespecified test on gpt-oss that its prespecified rule left undecided. The reversible-operator vocabulary does not apply to causal decoders, which receive their arrow from the mask; we treat decoders separately. We give no theory of the MLP beyond its role as the feature writer and the main currency of compensation. We have not yet evaluated protection-complete designs against qualified baselines; that sweep is specified but has not run (Section 8 and Appendix C). We do not claim that every successful sequence model must implement six named mechanisms, or that one architecture is globally minimal. The operator theory is developed in Chapters 1–4 of the monograph (Bruce, 2026a), which also holds every companion result this paper cites (Appendix A.1); the two-sided law is proved in Bruce (2026c), and the machines beyond the transformer that satisfy the lattice are mapped in Bruce (2026e).

2 The object: adaptive mixing in an anchored block

The operator theory used here is developed in Chapters 1–4 of the monograph (Bruce, 2026a); this section summarizes it, fixes notation, and records what the block adds to the operator.

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

$$P(X) = \mathrm{softmax}_{\mathrm{row}}\left(XW_Q W_K^\top X^\top/\sqrt{d_k}\right)$$

and delivers $P(X)XW_V W_O$. The kernel is row-stochastic, $P\mathbf{1} = \mathbf{1}$: one step of a random walk on the tokens (Bruce, 2026a, §1.1). 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 cannot amplify. With $D_\pi = \mathrm{diag}(\pi)$ and the stationary reversal $P^\dagger = D_\pi^{-1}P^\top D_\pi$, a head splits into a reversible part $S = (P + P^\dagger)/2$ and a circulation $J = (P - P^\dagger)/2$. J is skew-adjoint in $L^2(\pi)$ and is the only part that can route around a cycle (Bruce, 2026a, §3.2 and Theorem 10.1). A head with symmetric scores (queries and keys from one map, no mask) is reversible, $J = 0$ (Bruce, 2026a, §1.2). Reversal maps $S + tJ$ to $S - tJ$, so every quadratic readout $\langle Y, GY\rangle_\pi$ of one amplified head G is an even function of the circulation, although an output energy $\|GY\|_\pi^2$ need not be (Bruce, 2026a, Theorem 3.4). Depth, restarts and residual streams amplify the step by nonnegative combinations of its powers. Normalized, such an amplifier is $f(P) = \mathbb{E}[P^K]$, a walk of random length K whose law the architecture chooses; its anchor is the share $\Pr[K = 0]$ of the output that is the unpropagated input (Bruce, 2026a, §1.3). 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 is not used for them (Bruce, 2026a, §2.5).

Abstention. Off-by-one normalization, learned sink logits and value-null registers make the kernel substochastic on the real tokens. Each factors as $P = \mathrm{diag}(g)\,\mathrm{softmax}_{\mathrm{row}}(s)$, with gate $g_i = \mathrm{sigmoid}(\mathrm{LSE}_i(s) - b_i)$ and b_i the log of the null's exponential weight in row i; a headwise output gate supplies $g_i = \mathrm{sigmoid}(w^\top x_i)$

from the query instead (Bruce, 2026c). The softmax Jacobian $\mathrm{diag}(p) - pp^\top$ is positive *semi*definite, with null vector $\mathbf{1}$: the uniform logit shift, which would mean "attend less", does nothing. With a null it is positive definite, $\lambda_{\min} \geq (1-r)\min_i p_i$ at row mass r, and the potential $\log(e^b + \sum_j e^{s_j})$ stays convex (Proposition 14). The direction that lets a head abstain and the convexity that makes its logits easy to optimize live in the same potential.

Signed mixing. Differences of kernels, such as $P_1 - \lambda P_2$, have signed weights and rows summing to $1 - \lambda$. The constant mode is partly deflated, and outputs can leave the convex hull of the values. Multi-head attention already has this capacity, through signed value and output maps across heads (Bruce, 2026a, §15.5).

Adaptivity and feedback. Because $P = P(X)$ is rebuilt from its input, the Jacobian of $X \mapsto P(X)X$ in a direction U contains a feedback term $(\mathrm{D}P)[U]X$ that is sign-indefinite even when every kernel is positive and reversible. For the anchored iteration $X_{t+1} = \alpha P(X_t)X_t + \beta A$ with $0 \leq \alpha < 1$ and $\beta = 1 - \alpha$, contraction on a radius-R ball in the maximum-row norm is certified by the margin $m = \beta - \alpha\kappa_f > 0$, where $\kappa_f = RL_P$, with L_P the Lipschitz constant of $X \mapsto P(X)$ on the ball (the kernel measured by its largest absolute row sum), and $L_P \leq 2sR$ for dot-product softmax with scores $s\langle x_i, x_j\rangle$ at logit scale s; the bound needs the ball, since dot-product attention is not globally Lipschitz (Kim et al., 2021). The certificate is sharp over the class of smooth builders (Bruce, 2026a, Theorem 7.1). A sink brings the row mass into it: if every row carries mass at most $r_{\max} \leq 1$ and the weights have Lipschitz constant L_Q on the ball, the update contracts when $q = \alpha(r_{\max} + RL_Q) < 1$, and a sink-softmax head with the same scores and a state-independent null logit has $L_Q \leq 2|s|R\, r_{\max}$, so $q \leq \alpha r_{\max}(1 + 2|s|R^2)$ (Proposition 16). 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 (Bruce, 2026a, §1.4). The response of the equilibrium to its input need not be nonnegative or non-expansive. Amplification comes from the feedback, not from the average.

State. Recurrent mixers keep a matrix state; with a gated delta-rule write it evolves as $M_t = a_t M_{t-1}(I - b_t k_t k_t^\top) + b_t v_t k_t^\top$. With unit keys the write $M \mapsto M(I - bkk^\top)$ is non-expansive for $0 \leq b \leq 2$. Non-expansiveness bounds the difference between two states driven by the same inputs, not the state itself: without decay, unit keys whose directions converge while their values conflict can make the state grow like $\sqrt{t}$ (Bruce, 2026a, §18.2). A decay gate $a_t \leq a < 1$ contracts strictly and, with bounded values, bounds the state at every length (Section 6).

The block. A pre-normalized residual block adds 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).

3 Six protections

Each protection is a row of a conditional design inventory. It is stated for a declared class, with the paper that proves it, the signature observed when it is removed, and the evidence; Appendix B lists each quoted experiment with its evidence tier and the folder that holds its records. The rows do not say that every successful sequence model implements six named mechanisms. They say what goes wrong, and at what price, when a model in the stated class lacks the property.

P1 Anchoring. *Statement.* Let $f(P) = \mathbb{E}[P^K]$ be a normalized amplifier of a lazy reversible operator with content eigenvalue $\mu_C \in (0, 1)$. Collapse stays bounded as the filter is strengthened if and only if the path-length law keeps uniform short-path mass, $\Pr[K \leq m] \geq \rho$ for some fixed m and $\rho > 0$ (Bruce, 2026a, Theorem 2.8). The same criterion decides the retention of any interior positive mode of a fixed operator (Bruce, 2026a, 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 $2.3/(1 - \mu_C)$ layers when $1 - \mu_C$ 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

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]
re-injection	4	0	1.000	0.933	0.904	> 96
re-injection	4	1	1.000	0.972	0.959	> 96
re-injection	4	2	1.000	0.954	0.939	> 96
re-injection	*16*	*0*	*0.207*	*0.189*	*0.184*	*did not train*
re-injection	16	1	1.000	0.999	0.997	> 96
re-injection	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

Trained at $K_{\text{train}} = 64$: EMA off 0/6; EMA on 3/6.

Table 1: Looped models with and without input re-injection: one weight-tied Pre-LN causal attention plus MLP block on MQAR; 2,048 answers per value, chance 1/64. Conservative sequence-level 95% interval half-width is at most 0.063. EMA-of-latent starts at the input. Horizons are rounded with brackets; > 96 is right-censored. Untrained cells are italicized and excluded from horizon comparisons.

that mix, so a floor needs a bounded budget (Bruce, 2026a, Theorem 2.11). Without short paths, repeated averaging produces the rank collapse of pure attention (Dong et al., 2021) and the oversmoothing of graph networks (Oono & Suzuki, 2020); the path-law characterization is what this row adds. *Removal:* removing the identity path sends the minimal block to exact chance at depth 12 and beyond. In a 22M-parameter block, removing the attention sublayer's identity path gives exact chance, a constant output. Neither outcome isolates the collapse mechanism. The collapse horizon of plain graph-convolution operators is met on schedule; 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, shows a trainability cliff at depth ≥ 18 on CIFAR-10, bimodal across seeds. 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 1). *Compensation:* of the deficits tested so far, short-path mass is the one that nothing else in the block made up for, and under strict mixing no compensator of bounded gain can; at finite depth, large gain can (Bruce, 2026a, §5.5). Whether the horizon governs trained deep and looped transformers at scale has not yet been tested (Appendix C).

P2 Alignment (separation). *Statement.* Routing and content need independent maps when tied values would inherit the keys' nuisance, as exact tying does. Amplifying before or after a second map leaves one of two channels open, and for a reversible operator each has worst case equal to the input's norm times the spread of the gain over its spectrum (Bruce, 2026a, Theorem 4.1). A free value path closes the channel through which nuisance that rides in the features reaches the readout; Section 4 states when that is necessary and what it costs. *Removal:* a value map tied to the keys, $W_V = W_K$, fails a routing-and-content task at head dimension d_b (retrieval 0.29 against 0.97 separated, with value noise; means over 32 and 128 tokens), and the failure is dimensional rather than informational. *Compensation:* in the deeper models tested, MLP capacity absorbs the tying tax. At depth 24 and width 192 on CIFAR-100 the tax is +1.70, +0.04 and −0.72 pp at MLP ratios 0, 1 and 4 (five seeds), and at depth 12 it also vanishes by ratio 4; in four-layer models it persists (+0.65 and +0.78 pp at ratio 4, widths 96 and 192) (Bruce, 2026a, §6.6).

P3 Gain control. *Statement.* An anchored adaptive iteration contracts when the anchor share $\beta = 1 - \alpha$ outweighs the feedback, $m = \beta - \alpha\kappa_f > 0$, with $\kappa_f = RL_P \leq 2sR^2$ for dot-product softmax with logit scale

s on a radius-R ball; with row mass at most $r_{\max}$ the condition reads $q = \alpha(r_{\max} + RL_Q) < 1$ (Section 2; Bruce, 2026a, Theorem 7.1). The anchor, the logit scale and the representation radius trade against one another. The statement concerns the full feedback Jacobian: normalization bounds the radius but is not by itself a certificate. *Removal:* past the certified margin a smooth builder can have two attracting fixed points, and the certificate is sharp over that class (Bruce, 2026a, Theorem 7.1). At the largest response of a synthetic search, the response to the input grows from 1.77 with the weights frozen to 18.9 when they adapt, with zero circulation, both in the Euclidean norm (Bruce, 2026a, §12.2). Coupling the logit scale to the anchor has not yet been tested in trained language models (Section 8.5).

P4 Abstention. *Statement.* A head that can only route—rows summing to one over values it cannot zero—cannot abstain: it cannot lower its total mass or suppress its values independently of their contents. Its output is a convex combination of its values, zero only when they cancel, and the uniform logit shift that would mean "attend less" is the null direction of its Jacobian (Section 2). It can abstain only through a value-null option (a sink, a register, an off-by-one null) or by relaxing row mass (a gate, sigmoid attention). Every zero-valued null patch, and an output gate, factors as a gate times a softmax, while sigmoid attention lies outside that class (Bruce, 2026c). How the patch must scale with context is the two-sided law of Section 8. An abstaining head keeps a stochastic core: conditioned on long-run survival (a Doob h-transform), an irreducible sub-stochastic head is again an attention call, and so is its reversal (Proposition 17). *Removal and relaxation:* sigmoid attention uses the freedom (row mass 0.34 and 0.31 at depths 12 and 24) and cuts sink incidence from 0.53 to 0.22 at depth 12 and from 0.65 to 0.07 at depth 24, at a cost of 2.5 and 1.7 points of test accuracy (a vision transformer trained from scratch on CIFAR-100, three seeds). Output gating halves content-sink mass in a small hybrid causal character-level model with two softmax layers, over three seeds; the stationary-mass sink detector is degenerate on causal operators, so this readout carries that caveat. A causal mask makes the first token absorbing (Bruce, 2026a, Theorem 2.16), and trained models load it, as studies of attention sinks found (Xiao et al., 2024; Gu et al., 2025; Barbero et al., 2025). In a census of seven public causal language models at 64 tokens, 2,700 captured operators (2,544 heads and 156 head-averaged layers, each averaged over 16 inputs), the minimum row mass that each averaged operator sends to token 1, which is at least the mean of its per-input minima, has median 0.416 (0.410 per head). These are captured operators, not production-length measurements. *Abstention meets anchoring:* under a zero-value sink, a token with update share p, sink mass s and self-weight d leaves itself in a layer with probability $p(1-s-d)/(1-ps)$, which falls as s grows, so a sink is a learned per-token anchor dial (Bruce, 2026a, Proposition 8.6), close to the reading of sinks as a guard against over-mixing (Barbero et al., 2025). In sixteen pretrained language models the content-mixing budget grows as $L^{0.07}$ in Pythia, $L^{0.38}$ in GPT-2 and $L^{0.67}$ in OPT, and deeper Pythia models route their surplus attention into the sink, whose budget rises from 0.25 steps per token at 6 layers to 1.66 and 1.73 in the two 32-layer models (norm-based shares; each exponent is fitted across only four to seven models of one family; Bruce, 2026a, §2.7). The gain-control cost of dropping unit row sums at depth has not yet been measured (Appendix C).

P5 Routing budget. *Statement.* A single key competing with N others needs a logit margin that grows like $\log N$, as is known for retrieval (Nakanishi, 2025; Veličković et al., 2025); the simultaneous budget with a null is what this row adds. With the N competitors at a common logit, a share $1-\eta$ for the winner needs the margin $\log(N(1-\eta)/\eta)$, and content or noise in the keys adds to it (Theorem 4). With a null present the exact form is the simultaneous budget of Bruce (2026c): with content logits bounded by τ, no null offset keeps abstention a_* and match weight w_* beyond $N_* = e^{2\tau}(1-a_*)(1-w_*)/w_*$. At head width no larger than the content dimension, when content must be recovered through the keys, the budget is also dimensional (Section 4; Bruce, 2026a, §6.1). Along a sequence this budget is the price of routing by circulation: a circulating head can route within it, while a bidirectional circulation-free head that puts weight at least $1-\eta$ on each token's successor along n tokens needs scores spanning at least $\frac{1}{2}[(n-2)\log((1-\eta)/\eta) - \log n]$, which grows linearly in n when $\eta < 1/2$ (a corollary of Bruce, 2026a, Theorem 10.1(c), not yet checked in Lean). *Evidence:* on a synthetic retrieval substrate the accuracy onset grows as $\exp(2.14 \cdot \text{gap})$ in the trained logit gap, over 15 of 15 cells; on the resolution axis of vision transformers trained from scratch on CIFAR-100 and Tiny-ImageNet and tested on upsampled images, a training-free $\log N$ re-score recovers +0.65/+0.83 pp on average, 3–5% of a resolution deficit of 43 and 30 pp, though in a registered comparison a fitted constant temperature did at least as well (Bruce, 2026c), so this axis does not single out the $\log N$ shape, and building

the budget into multi-resolution training gains +6.3/+6.4 pp at the top resolution; in a prespecified test on trained ViT-S checkpoints, a $\log N$ re-score at evaluation raised accuracy on images upsampled to 16× the training tokens by +4.9 pp, and plain softmax gained at least as much as the tempered sink, so the gain comes from scaling the scores, not from the sink, and the same re-score raised accuracy by +4.5 to +4.9 pp on resized original images, most of them themselves upscaled at 16×. In trained causal language models with rotary positions, the routing budget is paid by position inside the trained range and by the positional treatment, interpolation such as YaRN, beyond it (Chen et al., 2023; Peng et al., 2024; Bruce, 2026c). The same law governs abstention (Section 8).

P6 The arrow. *Statement.* A single head routes around a cycle only through circulation. For a circulation-free head with a positive stationary law, the successor-to-predecessor odds around every cycle of tokens whose successor and predecessor probabilities are positive have geometric mean exactly one; such a head can swap tokens but not shift them, and on an open sequence it can point forward only by piling stationary mass exponentially at a boundary (Bruce, 2026a, Theorem 10.1). The statement 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 neighbours. On a synthetic cyclic-retrieval task (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 neighbour to the token itself (Bruce, 2026a, Theorem 10.1). Decoders take their arrow from the mask. 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 pp lower), and without a mask they are at parity on bidirectional lookup: a cost of the Gram form, not a removal of the arrow. *Reversal:* reversing learned circulation in trained retrieval models turns successor retrieval into predecessor retrieval, and in three pretrained vision transformers it changes predictions 8–12 times more than reversing feasible circulations of equal magnitude (Bruce, 2026a, §10.4). No statistic that is invariant under reversal can see the sign of these effects (Bruce, 2026a, Theorem 3.4).

Remark 1 (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 (Bruce, 2026a, §11.2). For a trained pure-transport mixer, dictionary retrieval R^2 is 0.18–0.26 against ≥ 0.99996 for additive blocks, and CIFAR-100 accuracy is 10 pp lower. Adding a feature-writing channel restores parity on CIFAR-100 classification but not on dictionary retrieval, where mean R^2 over five seeds stays at 0.19–0.23 at the training size of 16 associations. 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 W_O write, and the MLP is the one block component whose theory we leave open.

4 Structure-content separation: why the value path must be free

An architecture does not choose where gain sits relative to a second map; it chooses a parameterization, and training chooses the parameters. This section asks which parameterization lets training keep nuisance out of the content a head transports. It restates an earlier separation theorem in the two-channel terms of the monograph's Chapter 4, proves necessity only for an identified model, and replaces an unproved capacity claim by a restricted theorem. Proofs are in Appendix A.

Propagation architectures. 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 g, 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 2 (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). 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 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 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 $\|S - S_0\|_2 \leq e^{2a'} - 1$ between the normalized kernel of noisy data and that of its centers, with probability at least $1 - \delta$, simultaneously over every linear metric of bounded norm and every bandwidth above a floor, including ones selected from the same sample; a' grows with the mapped noise radius relative to the bandwidth, so the bound is informative only at high signal-to-noise ratio (Theorem 7). On four equally sized groups with centers $c_{yz} = y\, m_y u + z\, m_z v$, $y, z \in \{\pm 1\}$, and a metric that weights the class and nuisance directions by w_y and w_z, the ideal operator at bandwidth h has eigenvalues 1, $\tanh(w_y m_y^2/h^2)$, $\tanh(w_z m_z^2/h^2)$ and their product, on the constant, class, nuisance and interaction functions (Proposition 8). 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.

Applied to the normalized kernel, Theorem 3 needs a mild generalization that its proof gives. With equal group sizes the ideal kernel is stochastic, and its interaction eigenvalue, the product of the class and nuisance eigenvalues, lies in $[0, b]$; because the noisy kernel need not be stochastic, the constant direction enters the eigenvector bound, which then holds with $a - b - e$ replaced by $\min(a - b - e,\, 1 - a - e)$. 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. If the value map factors through the key map, $W_V = MW_K$, the head's output at token i is $M\sum_j P_{ij}k_j$: the head carries nothing the keys do not encode, and $\ker W_K \subseteq \ker W_V$. 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. Injective content recovery from the keys forces $r \geq d_b$; at $r = d_b$, E is invertible, so every nonzero query reads content. More generally, for a query span of dimension s, injective content storage that no query sees is possible if and only if $d_b \leq r - s$ (Proposition 9). 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 10).

A restricted capacity theorem. The routing budget 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 4 (Robust routing and value recovery)**.** *With $m = \Delta - 2Be_q - 2Q\nu$, the target's softmax share satisfies $p_* \geq \big(1 + (n-1)e^{-m/\tau}\big)^{-1}$. 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 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 11).

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 (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 rises $0.29 \to 0.96 \to 1.00$ from d_b to $2d_b$ to full width 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 7).

The two-channel reading. The earlier account read a separated head as the architectural form of "project, then amplify". The monograph 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 (Bruce, 2026a, 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 above; later maps, and the anchor, which caps gain ratios on a lazy reversible operator (Bruce, 2026a, Corollary 4.2), are what act on those.

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 (P6), 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 (P1).

5 The lattice

We use 'lattice' for the Boolean lattice of subsets of relaxed constraints, ordered by inclusion, with the vertex at the empty set. Tables 2–4 list its generators, one constraint per row; not every element is a useful design. The "vertex" marks where softmax attention in a pre-normalized residual block sits. The columns give what relaxing the constraint buys and what must then be re-supplied (for the block protections of Table 4, the signature of removing one and whether its loss is compensable), and where practice sits; the last column names the registered tests that bear on the row, each described in Appendix C. Figure 1 draws an illustrative slice of the lattice, rows A1 and A2. 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.

Table 2: Kernel constraints.

Constraint (vertex)	Relax → buys	Must re-supply	In practice	Open tests
A1 Unit row sums (satisfied)	Abstention; strict contraction $\|P\|_w = \rho < 1$ for an irreducible head (Proposition 17); 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	Abstention harness at 350M and at depth
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	Head-count sweep
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	Circulation gates; reversible composition; diffusion LMs
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	—

Table 3: Dynamical and context constraints.

Constraint (vertex)	Relax → buys	Must re-supply	In practice	Open tests
B1 Frozen kernel (relaxed: adaptive)	Content-dependent routing	Gain control, a positive margin $\beta - \alpha\kappa_f$ (with row mass: $q < 1$): anchor, bounded logit scale, bounded radius	Residual, LayerNorm, QK-norm, soft-capping (z-loss acts on the output logits)	Coupled logit scale

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	Hybrid ratio sweep
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; on top of YaRN's own temperature it raised long-context perplexity, as YaRN's temperature applied without interpolation also did	$\log n$ scaling, YaRN temperature, scalable softmax	Abstention harness at 350M

Table 4: Block protections.

Protection (vertex)	Remove → signature	Compensable?	In practice	Open tests
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	Anchor horizon at scale
C2 Alignment (W_V)	Tied value map fails at constrained head dimension	Yes (MLP capacity, width)	W_V; latent KV	Compensation at scale
C3 Norm and gain control (LN)	Trainability cliffs; logit growth	Partly	Pre-LN, QK-norm, hybrid norm	Coupled logit scale; abstention harness at depth
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	Head-count sweep

Notes on the rows. *A1.* Abstention requires a way to reduce a head's update in the declared routing class, not a particular named sink; the gate-times-softmax factorization and the budget are those of Bruce (2026c). *A2 and C5.* Differences of heads already realize subtraction inside multi-head attention, so a signed head adds capacity mainly when heads are scarce (Bruce, 2026a, §15.5). *A3.* The arrow statement is per head (P6). *A4.* Row-normalizing a symmetric affinity gives a reversible walk (Bruce, 2026a, §1.2), and tied positive-semidefinite scores give a stochastic kernel whose spectrum lies in $[0, 1]$ (Proposition 15); indefinite scores buy negative eigenvalues. *A5.* Exact sparsity is analysed in Bruce (2026a, §6.4).

B1. The margin is the certificate of Section 2; the logit scale, the radius and the anchor trade against one another (Bruce, 2026a, Theorem 7.1 and §7.3). *B2.* Non-expansive writes alone bound no state; a decay gate $a < 1$ does for bounded keys and values, 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 (Bruce, 2026a, §18.4). In the length comparison, whose recurrent arms were non-expansive but not certified contracting, the exploding arm was the diagonal state-space model with $|a| = 1$ (Section 8). 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 (Proposition 12), whereas additive writes make the affine updates non-commuting

($x \mapsto 0.5x + 1$ against $x \mapsto 0.9x$). The group-theoretic statement is due to Grazzi et al. (2025) and Siems et al. (2025). *B3.* With content logits bounded by τ, no null offset, fixed, learned or shifted by $\log N$, keeps both sides beyond $N_* \propto e^{2\tau}$; tempering the whole budget keeps both when the margins are strictly separated, which no head of the trained 125M language models met (Bruce, 2026c).

C1. The price law is the path-law calculus of Bruce (2026a, Chapter 2). The gain on a content mode μ_C is $f(\mu_C) = \mathbb{E}[\mu_C^K]$, so $\kappa = 1/f(\mu_C)$. A family keeps a floor if and only if it keeps a fixed share of path mass on a bounded window (its Theorem 2.1). An identity atom suffices but is not necessary in discrete depth, and it is necessary in continuous depth (Proposition 2.3). With per-layer gains bounded below, the rate is $\exp(-\Theta(S_L))$ in the expected number of mixing layers (Theorem 2.11 and Corollary 2.12): the tenfold horizon is $S_L \approx 2.3/(1-\mu_C)$, or $L^* \approx 2.3/(1-\mu_C)$ for plain powers. The vertex needs qualifying. A fixed mixing share p only delays collapse, to about $2.3/(p(1-\mu_C))$ layers; a floor needs a bounded budget, as restarts, a fixed input atom or depth-scaled branches supply. At initialization Pre-LN is critical: with s the attention share of branch energy, its budget is $s \log L$ and, for incoherent updates, relative modal energy falls as $L^{-s(1-\mu^2)}$ (Bruce, 2026a, Proposition 2.21). In sixteen pretrained language models, with norm-based shares, the content-mixing budget grows much more slowly than depth, as $L^{0.07}$, $L^{0.38}$ and $L^{0.67}$ in Pythia, GPT-2 and OPT, and it is nearly depth-independent only in Pythia (Bruce, 2026a, §2.7). *C2.* Three attempts to make the value channel compact or independent did not open room. 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 2× bar in every seed (at the first layer, already near exact, it cleared the bar); a rank-zero arm already solved the task; and, 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. *C3.* LayerNorm is a norm-control mechanism, not a certified cap: its Lipschitz constant is of order $\epsilon^{-1/2}$, where $\epsilon > 0$ is its variance stabilizer. *C4.* The transport-only statements hold for the pure transport mixer, whose Gram invariant the writing variants break (Remark 1; Bruce, 2026a, §11.4).

Within a declared class, content costs one of three currencies. 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 (Bruce, 2026a, Theorem 2.1). Leaving that class buys content in one of two other currencies: signed coefficients, which leave the stochastic cone (Bruce, 2026a, Proposition 1.4 and §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: $P_L = (1 - L^{-2})I + L^{-2}U$ with $U = \mathbf{1}\mathbf{1}^\top/N$ keeps centered gain $(1 - L^{-2})^L \to 1$ at depth L, because its operator changes with L, outside the fixed-mode premise; external gain or a new source of information changes the model as well.

6 A successor class retraces the lattice: recurrent and streaming mixers

Recurrent and streaming mixers are the main alternative to attention, and they meet the same failure modes with their own mechanisms. Each protection of Section 3 reappears under another name.

The inventory retraced. A decay gate is the anchor. A delta-rule write with step $\eta > 0$ and decay $\lambda \geq 0$, $M \mapsto M\big(I - \eta(\lambda I + kk^\top)\big) + \eta v k^\top$, is non-expansive in the spectral and the Frobenius norm when $\eta(\lambda + \|k\|^2) \leq 2$, but non-expansiveness bounds only differences between states written with the same keys and values, not the state itself: without decay, unit keys whose directions converge while their values conflict can make the state grow like $\sqrt{t}$ (Bruce, 2026a, §18.2). Strict contraction bounds the state at every length. With keys of norm at most K, values of norm at most V and $q = \max(|1 - \eta\lambda|, |1 - \eta(\lambda + K^2)|) < 1$, $\|M_t\| \leq q^t \|M_0\| + \eta V K (1 - q^t)/(1 - q)$, and with bounded readout logits the per-token loss is bounded too; the write condition is the class's gain control. 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. A fixed-size state pays for its length robustness in recall. 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 (Bruce, 2026a, §18.3). Hybrids keep a few attention layers as the routing-budget reserve for recall, and output gates and denominator sinks carry abstention here as they do in attention.

In-system evidence. The write certificate is consistent with the observed stability outcomes in small trained fast-weight memories. 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 63 cells with momentum that stayed bounded between the certificate and the measured stability edge count as matches, and in 74% when they do not. A causal test inside a trained model is still to run. A non-expansive flow block degraded 1.24–1.56 times in bits per byte at eight times its training context, against 2.18–4.22 times for transformers with sinusoidal or no positional encoding (in-sample and averaged over every position of that context, two scales, two seeds) (Bruce, 2026a, §18.6). That robustness costs 0.29 and 0.33 bits per byte at the training length against the sinusoidal transformers (2.61 against 2.32 and 1.98 against 1.65), and no rotary or ALiBi transformer was compared (Su et al., 2021; Press et al., 2022). The decay gate acts as the anchor inside a trained hybrid. 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; that run's records are outside this package). 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 numbers come from that experiment's records, reconciled between its two runs; the software environment of the original runs is not verified. Under one frozen matched-data protocol, 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 (Bruce, 2026a, §18.3). That protocol's attention baseline is not yet qualified, so no ranking between the two is claimed. Latent recursion shows the anchor directly. A weight-tied block trained at 4 loops keeps 0.904–0.959 accuracy at 96 loops when its input is re-injected and 0.023–0.032 when it is not (three seeds); trained at 16 loops it keeps 0.994–0.997 in the two re-injection seeds that trained (the third did not), against 0.183–0.228. An exponential moving average of the latent, whose state starts at the input, also delays collapse: 0.835–0.884 at 96 loops when trained at 16, still falling from 0.983–0.989 at 64 (Table 1).

The causal resolvent as a streaming primitive. The anchored aggregator $Y = \beta(I-\alpha P)^{-1}X$ of a frozen causal kernel needs no matrix inverse. Row i of $(I - \alpha P)Y = \beta X$ gives $y_i = \big(\beta x_i + \alpha \sum_{j<i} p_{ij} y_j\big)/(1 - \alpha p_{ii})$, and every denominator is at least β (Proposition 13). For a finitely factorized kernel, $p_{ij} = u_i^\top v_j / z_i$ with $u_i, v_i \in \mathbb{R}^r$, the prefix sum is $u_i^\top H_{i-1}/z_i$ with $H_i = H_{i-1} + v_i y_i^\top$: one forward pass with $O(rd)$ state, independent of the sequence length, gives an anchored aggregator at linear-attention cost. The constant memory comes from the factorization and does not extend to dense softmax, whose prefix can require every earlier output; and a kernel rebuilt from its own output turns the recursion into an implicit system that this argument does not cover.

7 Why the transformer is a fixed point

Two measured mechanisms help hold the design where it is. Necessity is real but local, and composition buys it back (compensation). Protections made into explicit knobs do not beat a block that already pays for them

(mechanisms are not knobs). A third is a hypothesis: if prices are concave, optimal allocations sit at corners, so designs would jump rather than drift (corners). A blueprint reading and the retrodiction table follow.

Compensation. Necessity theorems bound what one constrained map does in one application, and they are exact on their own constructions: for $C = 10$ classes, a rank-$(C-2)$ content map on the proof's construction sits at 90.16% (five seeds, 89.2–91.1%), at the ceiling $(C-1)/C = 90\%$ for a merged class pair; the seeds scatter around it because the merged pair stays 0.026 apart and the test classes differ in size. Trained systems evade them by composition wherever another resource has room (Bruce, 2026a, §15.1). 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. At depth 24 and width 192 on CIFAR-100, MLP capacity turns the value-tying tax from +1.70 to −0.72 pp as the MLP ratio grows from 0 to 4, while in four-layer models the tax persists at ratio 4 (Section 3). Heads buy back subtraction: softmax difference retrieval rises from $R^2 = 0.46$ with one head to 0.998 with four. The neighbourhood of the standard design is therefore flat where depth gives the MLP room to compensate; in shallow models the tying tax remains. 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 (Bruce, 2026a, §5.5). Whether the compensation budget holds at larger scale has not yet been tested (Appendix C).

An explicit knob does not beat the block's own mechanism. 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 (Bruce, 2026a, §15.4). The mechanism layer held in several rows, 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 every row 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. Standard components already pay for the protections. The residual is the anchor; LayerNorm is norm control, although not a certified cap; and multi-head attention with signed value and output maps subtracts (Bruce, 2026a, §15.5).

The corner hypothesis. If prices are concave, allocations of protection strength sit at corners. Under concave prices an optimum over the declared feasible set is attained at a vertex, and every optimum is a vertex when the prices are strictly concave; a vertex has at most one fractional coordinate, and greedy ordering can fail under capped anchored pricing, so the allocator enumerates corners (Bruce, 2026a, §16.2). The species prices behind it are those of Bruce (2026a, §1.3). The anchored species' premium is robustness at a fixed strength t: the restart's log-price sensitivity to the content gap $\beta = 1 - \mu_C$ 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 the restart's own collapse ratio, $\kappa_{\text{res}} = 1/f(\mu_C) = 1 + t\beta$. The clause "at fixed strength" matters. At matched suppression $1/x$ of a nuisance gap $\beta_n > \beta$ 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 $\beta/\beta_n = 0.2$ (direct differentiation; Bruce, 2026a, §16.1). 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 (Qwen Team, 2025; Kimi Team, 2025). This record is descriptive and was curated by the hypothesis holder. On a single attention-fraction axis, concave pricing predicts 0, 1 or a cap, not a particular fraction, and the level set $\{1/2, 1/3, 1/4, 1/6, 1/8\}$ named in the corner prediction (Section 9) is a convention, not a consequence (Bruce, 2026a, §16.3). 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 that Bruce (2026a, §16.4) sets out.

7.1 From map to blueprint

Run forward, the inventory is a constraint set, and the smallest block that satisfies it can be assembled and tested. The inventory asks for an independent content map (Section 4), short-path mass (P1), a $\log N$ routing budget (P5), an abstention channel (P4), gain control (P3) and, outside the closed-form inventory, interleaved nonlinear writing (Remark 1). The minimal certified block (Bruce, 2026a, §16.6) 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 (P6), and a $\log N$ logit scale. In trained 125M-parameter causal language models such a scale helped beyond the training length on interpolated positions; on top of YaRN's own temperature it raised long-context perplexity, as YaRN's temperature applied without interpolation also did (Bruce, 2026c). Symmetric scores make every layer reversible by construction, and detailed balance holds to 10^{-12} on every layer of all 50 trained minimal encoders. 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, by +1.8 to +3.9 pp on CIFAR-10 and CIFAR-100, averaged over the five seeds of each dataset and learning rate, under one shared recipe tuned for neither model; at each model's better learning rate the margin is +2.4 pp on CIFAR-10 and +3.0 pp on CIFAR-100. This is tiny-scale evidence, not a scaling claim; the comparison at 22M parameters waits for tuned baselines (Appendix C).

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. 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, 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. (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 verdict 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 it held in none of the three tests. A dial fed a regressed estimate of the best strength lost to the best fixed strength, one fed a measured estimate fell short of it within a tolerance set in advance, and the certified allocator still beat the best fixed configuration. (d) Signed, learnable gain shaping helps where task relevance sits at low or negative eigenvalues, as on heterophilic graphs (Bruce, 2026a, §2.3).

7.2 Retrodiction

Table 5 reads proposed and adopted changes as lattice moves. The first row has an exact counterpart in the mean-field model at initialization. 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 (GPT-2's initialization, 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 (Bruce, 2026a, Proposition 2.21 and §2.6). The Post-LN cliff of P1 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. Section 9 states the predictions that can.

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 (Su et al., 2021) enter through row A3, where untied queries and keys let them circulate,

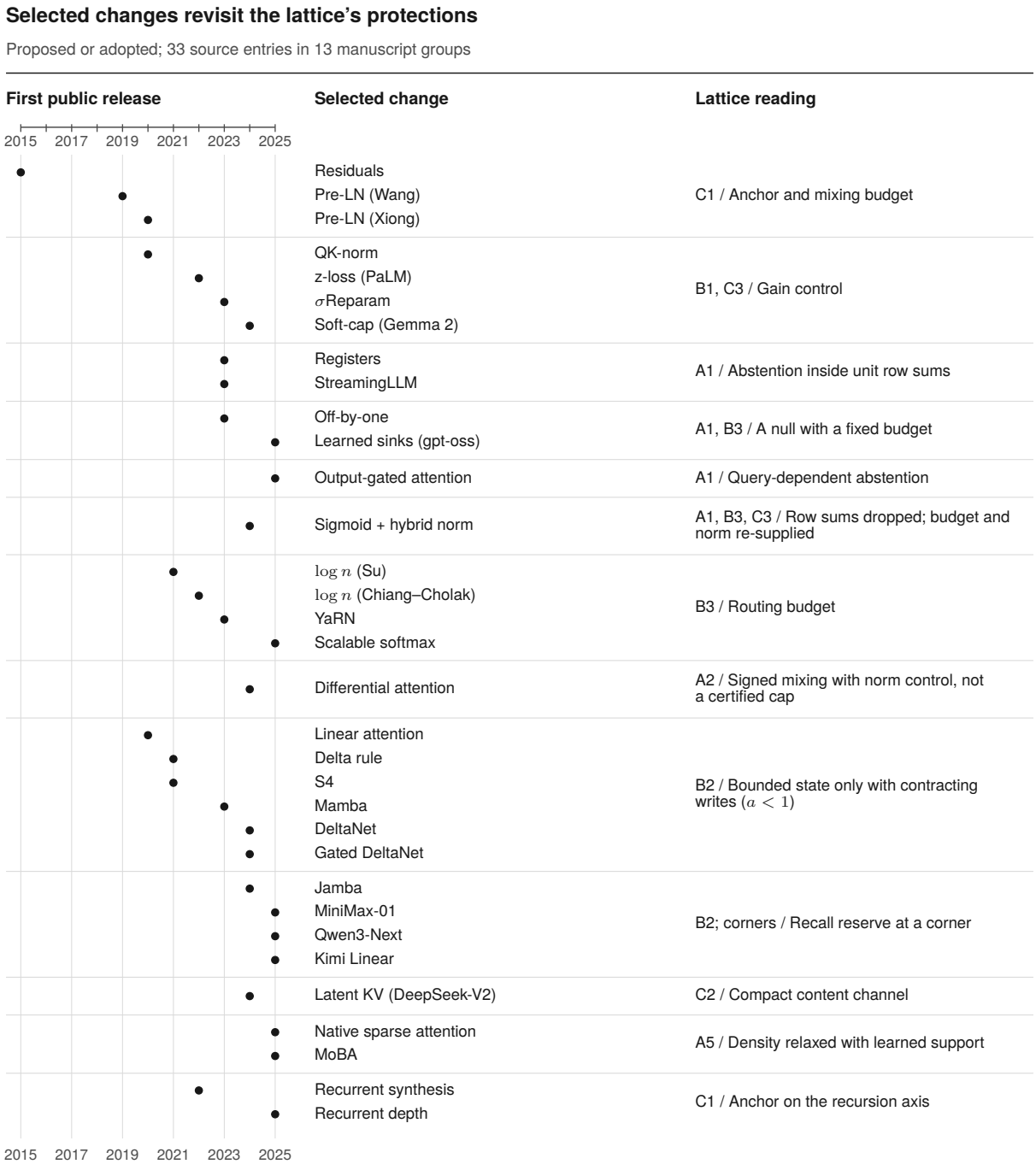


Figure 2: Selected proposed or adopted changes map to recurring protections in a retrospective reading. Each dot gives the first-public year of one cited source (arXiv v1, blog or model-card release), not an adoption date. The 33 entries split the first 13 of the 14 groups of Table 5, without Yang et al. (2024a); vertical position within a group separates sources and encodes no quantity. Row labels refer to Tables 2–4. 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 retained in the figure provenance.

Table 5: 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. Readings are interpretations, not evidence for the lattice (Appendix B).

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 at initialization (trained Pre-LN models measure $L^{0.07}$ to $L^{0.67}$)	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 $\kappa_f \leq 2sR^2$ (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 8)	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., 2024b; 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., 2024a; 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

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.

8 Where the room is

This section keeps to where the transformer's own protections are patched, overpaid or unpaid. Machines beyond the transformer, including its recurrent successor class, are mapped in Bruce (2026e).

8.1 The two-sided log-N law of abstention and retrieval

The law is proved in Bruce (2026c); we state the parts the lattice uses. A head sees one matching key with logit s_* on present queries, N background keys with logits s_j, and a null with weight $z_N = e^b$. Write $B_N = \sum_j e^{s_j}$. The abstention mass on an absent query is $a_N = z_N/(z_N + B_N)$, and the match weight on a present query is $w_N = e^{s_*}/(e^{s_*} + z_N + B_N)$.

Theorem 5 (Two-sided law; Bruce, 2026c)**.** *(i)* Exact budget. *Both $a_N \geq a_*$ and $w_N \geq w_*$ hold if and only if $a_* B_N/(1-a_*) \leq z_N \leq e^{s_*}(1-w_*)/w_* - B_N$. If every content logit lies in $[-\tau, \tau]$, no choice of null, fixed, learned, or shifted by $\log N$, satisfies both beyond*

$$N_* = e^{2\tau}(1-a_*)\frac{1-w_*}{w_*},$$

and the bound is sharp. (ii) Fixed and calibrated nulls. *A fixed null gives $a_N \leq \mathrm{sigmoid}(b+\tau-\log N)$, and its match weight also falls, $w_N \leq e^{2\tau}/(e^{2\tau}+N)$; retrieval survives only as accuracy under a decoder. A null raised by $\log N$ keeps abstention and loses the match weight once $s_* - b$ falls below $\log N$ plus a margin. (iii)* Tempering. *Multiply every logit and the null by $\lambda_N = \log N/\log N_0$. If the background logits are at most $s_0 < b < s_*$ and $v_a = N_0 e^{s_0 - b}$, $v_r = N_0 e^{s_0 - s_*}$ and $u_r = e^{b-s_*}$ are at most one, then $a_N \geq (1+v_a^{\lambda_N})^{-1}$ and $w_N \geq (1+u_r^{\lambda_N}+v_r^{\lambda_N})^{-1}$ for every $N \geq N_0$, and both tend to one when all three are below one. (iv)* Converse. *For i.i.d. background logits with essential supremum s_0 and $h = \log N_0 + s_0$, tempered abstention tends to 0 if $b < h$ and to 1 if $b > h$.*

Without separated margins, tempering does not help: with $b = s_0$ and every background logit at s_0, the abstention is $1/(N+1)$ at every temperature. The theorem bounds attention weights under a fixed score law. It does not identify the score law a network learns, and it does not turn an accuracy onset into a weight threshold.

Controlled pilot. Mechanism-level lookup-or-abstain heads were trained at $N_0 \in \{64, 512\}$ with logit bounds $\tau \leq 10$ or 20 and evaluated to $N = 65{,}536$ keys. At $\tau \leq 10$ the fixed learned sink abstained, in seed means, at 0.87 at the training count $N = N_0 = 64$, 0.10 at $N = 4{,}096$ and 0.007 at $N = 65{,}536$, matching the mean-field law, evaluated with each head's learned logits, to within 0.002 per seed, while its retrieval accuracy stayed perfect. The length-calibrated sink and sigmoid with $-\log N$ held abstention flat (0.87 and 0.89) but lost retrieval from $N = 2{,}048$. In this geometry the $\log N$-tempered rule kept both to 65,536; in seven further settings it lost a side with heavy-tailed keys, and at the smaller budget with clustered keys and with four heads. A larger logit budget moved the failures of the rules that use it to larger N; for heads trained at $N_0 = 64$, the softmax and sigmoid onsets did not move with the logit bound (Bruce, 2026c).

Production model. gpt-oss-20b has learned sink logits in dense layers and in 128-token banded layers, whose background stops growing at 128 keys: the same mechanism at different key counts, a built-in control. A prespecified test of the sink length law there was undecided by its prespecified rule; descriptively, its slopes leaned toward compensation through position (Bruce, 2026c). Implementation and fidelity checks passed, but the books shuffle control's gap was 0.672 nats, below the required 1 nat, and the natural books text itself scored 7.06 nats per token (the code, 3.39), far above what such a model attains on ordinary prose, so the failed control may reflect the input rather than the model. The slopes of the sink log-odds on $\log(i+1)$ matched neither prespecified branch either. They were -0.20 in dense layers beyond position 256 and -0.22 while the band fills, where the prespecified criterion was at most -0.5, and -0.02 in banded layers beyond their window. The test neither confirms the law in a production model nor shows that the model compensates. It is no evidence against the identity either, whose slope of -1 assumes a stationary background. The claims about trained models rest on the controlled pilot and on the trained harness below.

Trained models. The trained harness 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 as the registered primary positional treatment (Bruce, 2026c). Its lead prediction, that the two tempered rules would degrade least with length, failed. On the same checkpoints, tempering on top of YaRN raised long-context degradation against softmax: the perplexity ratio at 8 times the training length, relative to the same checkpoint's perplexity on the same books' training-length prefixes, 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 which tempering itself, against the same checkpoint untempered, added 0.10 to 0.17 in each seed, and tempering did not hold abstention within the registered bound. None of the probe's retrieval heads met the separated margins the guarantee needs, so it was not exercised. Separating YaRN's two parts on the same checkpoints shows why: YaRN already carries a log-length temperature, so the tempered arms were over-sharpened, while the same schedule

on interpolated positions lowered the loss beyond the training length in every checkpoint tested (Bruce, 2026c). The comparison registered from this inventory held in these language models for the output gate and the tempered gated sink, and, descriptively, 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%, while a signed gate without a null did not (a 7% mean cut; its paired differences from softmax change sign across the three seeds). These reproduce, at matched training across the rules, earlier reports for heads that can do nothing, attention biases and output gates (Bondarenko et al., 2023; Sun et al., 2024; Qiu et al., 2025); what the harness adds is where the null lives once the length changes, below. Separately, sigmoid aside, every rule with a null lowered perplexity at the training length, by 1.2 to 2.5 percent, and the signed gate without a null lowered it too, by less. At this scale no retrieval head reached the registered threshold and no rule abstained behaviourally, so these abstention results describe attention weights, not behaviour. On text, fixed sinks compensated through position: a learned sink's log-odds fell with a median head slope, averaged over three seeds, of only -0.12 on $\log(i+1)$ inside the training length on FineWeb and -0.28 beyond it with YaRN, against -1 for a stationary background, but about -6.2 natively beyond it, and the heads probed for abstention did not compensate. Where the null lives matters once the length changes: natively, a fixed null in the scores collapses past the training length, while plain softmax, which has no null, and an output gate, whose null lies outside the scores, degrade less, and interpolated positions carry the rescue for all of them. Controlled byte-level models show the same ordering, and show that a background's growth adapts to the null rule only through position (Bruce, 2026c). Where the null lives never changes the arrow: in a head without a mask, a null or an output gate neither creates nor removes circulation, because conditioned on long-run survival the head is softmax with an added key bias (Proposition 17), and a key bias leaves the cycle part of the scores, the only part that circulates, unchanged (Bruce, 2026a, Theorem 3.2). The 350M leg has not run (Appendix C).

In the harness's vision transformers, the one registered prediction, that output gates lower both content-sink mass and massive activations below softmax's in every seed, failed: content-sink mass fell in all three seeds, the activation ratio in only two. Descriptively, tempering the same trained ViTs by $\log N$ at evaluation left the training resolution unchanged and raised accuracy at every higher resolution, on upsampled images with interpolated positions, the more so as the token count grew (Bruce, 2026c). A prespecified confirmation on the same checkpoints then found the gain at 16 times the training token count in every seed, and found that tempering plain softmax's scores gains as much. The benefit is log-N scaling of the scores, and the sink is not needed for it, which is not a claim that sinks have no other use (Bruce, 2026c). A second prespecified re-evaluation of the same checkpoints found the gain on resized original images as well, $+4.5$ to $+4.9$ points at 16 times the training token count, although at that size most originals are themselves upscaled. It also measured the attention rows. At the new resolutions the background grows in proportion to the number of tokens while the top score rises only slowly, and tempering moves the rows' attention further from its trained 224 px pattern rather than restoring it, so why the scaling helps vision transformers is open. Descriptively, the same rows show where the null lives in vision too: the learned sink's log-odds fall at about -0.96 per e-fold of the token count, close to the stationary law's -1, while with an output gate added to the sink abstention holds level (Bruce, 2026c).

A law of sums. The two-sided law is a law of sums. 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. In a trained model that condition belongs to the positions, not to the rule. With YaRN, the max-statistic head 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 means over 3 seeds, though not in every seed; 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 extrapolated positions raised the background maximum with length, the evidence wall the theory names, and its rescue is the interpolation's (Bruce, 2026c).

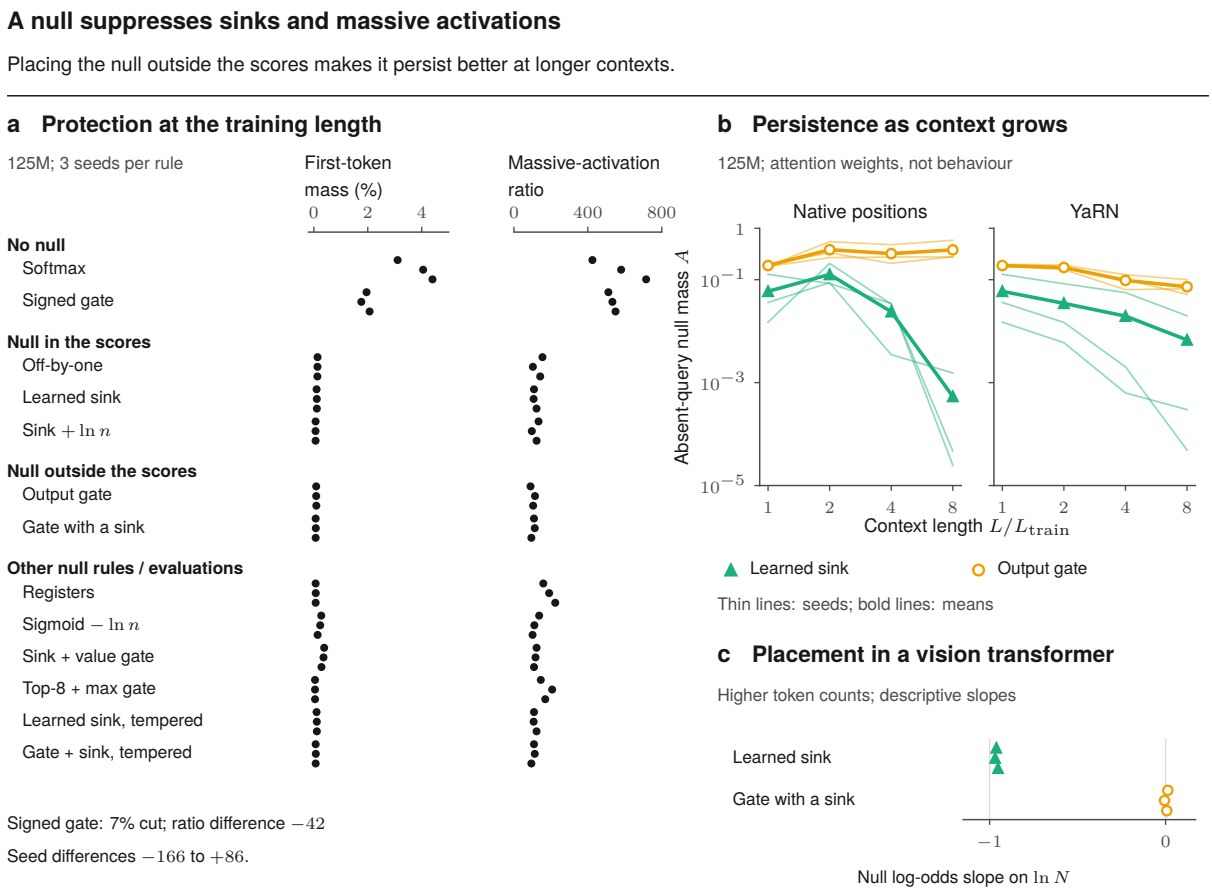


Figure 3: A null suppresses first-token sinks and massive activations in these language models, and its placement governs persistence at longer contexts. (a) All 13 evaluation variants of the 125M harness; each dot is one of 3 seeds, trained at 1,024 tokens. First-token mass is measured on FineWeb; the 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.

8.2 Protection-complete hybrids

A hybrid that keeps a few attention layers beside a certified contracting recurrence ($a < 1$) is the lattice's candidate for a resource-axis win. The test that would decide it, an attention-fraction sweep with gated attention under the registered positional treatment that measures quality at parity, degradation at 8×, state size and the certificate's contribution, has not run (Appendix C). 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. These are in-sample measurements on WikiText-103, with the loss averaged over every position of the longer context and the ratios averaged over two seeds at each of two scales; the largest model has 19.2M parameters (Bruce, 2026a, §18.6). The recall cost of bounded state is not established. With its curriculum on, the minimal arm of the recall test is bimodal across seeds (with it off, both seeds stay low), and it was trained at a preregistered learning rate not retuned for it, while a gated delta-rule recurrence recalls 32 stored pairs at 0.998–0.999, within its training load (five seeds; Section 6). It becomes a claim only after the recall-capacity sweep inside the hybrid experiment.

8.3 Direction as a knob

Two registered tests bear on direction as a design variable: whether a model can use an explicit per-head direction knob, a gate on each head's circulation for which the first-order predictor should anticipate the sign of the learned change, and whether two reversible layers learn to compose a directed block. Neither has run (Appendix C).

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. 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 we follow is never to 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.

8.4 Organized blocks

Blocks whose preference, exchange, circulation, abstention and anchor are explicit parameters have their construction identities checked to machine precision before training (Appendix B), roughly match a parameter-matched transformer at the training length on three small synthetic causal tasks, and, in the same in-sample comparison on WikiText-103, degrade less at eight times their training context than transformers with sinusoidal or no positional encoding (Bruce, 2026a, §18.6). Whether the minimal anchored block keeps its margin against tuned baselines, and which factor carries it, has not yet been tested (Appendix C).

8.5 Stability by construction

Gain control could be built into training rather than patched afterwards. If decoupled weight decay is modelled as the anchor, $\beta = \eta\lambda$, and the feedback weight as $\alpha = \eta$, the certificate's threshold is $s_c = \lambda/(2R^2)$: at a fixed domain radius R it is set by the weight decay alone, and the learning rate cancels (prediction 7; Bruce, 2026a, Theorem 7.1 and §7.3). This reads the AdamW step $W \mapsto (1 - \eta\lambda)W - \eta\nabla L$ as an anchored iteration whose feedback has Lipschitz constant $2sR^2$, an analogy that the registered 125M test is designed to check. A spectral cap on the query–key maps tied to λ should then remove this class of instability at matched quality, where QK-norm, soft-capping and σReparam control the attention-logit scale, whose growth small-scale proxies reproduce (Wortsman et al., 2023), and z-loss regularizes the output logits, all without reference to the anchor. No experiment has yet shown that the coupling removes this instability. The one test so far did not measure feedback instability: in its toy loop the instability flags tracked the anchor's relaxation, and in its byte language model the detector took normal attention sharpening for instability, so neither is

cited for gain control (Appendix B). A comparison of the coupled cap with the four standard fixes in a 125M language model, measuring the loop gain directly, is registered and has not run (Appendix C).

8.6 What did not work

Pure transport without writing; per-instance measured anchor dials, an honest tie once priced; engineered null registers, which cut mean content-sink mass by 14% in an eight-layer byte language model (five seeds), with no detectable advantage over a plain register token, on the caveated causal sink readout; eviction of cache entries by absorbed mass, which loses to heuristic eviction when judged in a linear model built from captured attention; free signed normalization, which did not win on heterophilous graphs and hurt homophilous ones; and compact value-channel designs, which saved cache but gained neither task performance nor decoding speed (row C2). 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 that survives only in a laboratory log. And in an unregistered comparison on CIFAR-10 (eight layers, width 128, three seeds), the free value map added 3.99 pp of accuracy to attention, while swapping attention for the resolvent, with the value map removed from both, changed it by -0.06 pp. The lattice anticipates most of these: pure transport cannot write (row C4); abstention needs only a way to reduce the update, not a particular sink (A1); signing buys subtraction, which homophilous graphs do not need (A2), although the heterophily win A2 would favour did not appear; the value map carries the gain (C2); and circulation rotates without amplifying (A3). Eviction by absorbed mass is a limit of the tested policy, which no row addresses.

9 Predictions stated in advance

Three of the predictions below have been tested: the sink length law, on a production model and in the trained 125M harness, where fixed sinks took the registered compensated branch; sigmoid retrieval, in the same harness; and, provisionally, transport-only mixers. The others have not yet been tested. Each is falsifiable with public models or small controlled runs, and Appendix C describes the experiments specified to test several of them.

1. *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 (Section 8); in the trained 125M harness the fixed sinks took the registered compensated branch. Since the slope is $-\gamma$, where γ is the measured growth of the head's background, the prediction is about γ: 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 with native positions; and lower in more local heads, shallower slopes tending to come with greater recency. A rotary head whose background grows at the stationary 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.

2. *Norm control accompanies dropped row sums.* Any widely adopted attention that drops unit row sums ships with a norm-control mechanism. 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.

3. *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 $[-\tau, \tau]$ and one offset shared by matched queries, fixed positive abstention and match weight cannot both hold beyond $N \propto e^{2\tau}$ (Bruce, 2026c). 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, about half the $\ln 8 \approx 2.08$ of a strict $1/N$ law; no head met the retrieval-identification threshold, so this concerns attention weights (Bruce, 2026c).

4. *Signed gains shrink with heads.* Gains from signed or differential attention shrink as the number of heads grows at fixed width.

5. *Circulation in diffusion language models.* Masked diffusion language models carry more delivered circulation in early layers than masked-LM encoders of similar size.

6. *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 (Bruce, 2026a, Theorem 3.4). The prediction concerns statistics not yet proposed: any future statistic that predicts the sign of reversing a trained head's circulation will change under reversal.

7. *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 (Bruce, 2026a, Theorem 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 (Bruce, 2026a, §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.

8. *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; the level set is a convention for reading the prediction, not a consequence of the theory (Bruce, 2026a, §16.3).

9. *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; 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 Bruce (2026a, 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 $\lambda_s \lesssim N^{-1/2}$. 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 $1/((L+1)(1-\mu))$; that law need not carry over to a Pre-LN residual loop. The shares p, λ_s and r are measured, not fitted.

10. *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 prediction concerns pure transport; variants that re-center or write break its invariant (Bruce, 2026a, §11.4).

The trained-harness 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 Bruce, 2026c 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, which already carries a log-length temperature, the rule did not keep abstention and raised long-context perplexity, so the outcome lies outside all four, and no head was certified (Bruce, 2026c).

10 Discussion and limits

A measurement principle. Default statistics are often invariant under a natural involution, and then they are blind to the odd component, which carries direction or contrast. For attention the involution is stationary reversal: every stationary-weighted quadratic readout $\langle Y, GY\rangle_\pi$ of one amplified operator G is

even in the circulation, so it can measure how much circulation there is but never its sign, although an output energy $\|GY\|_\pi^2$ need not be (Bruce, 2026a, Theorem 3.4). For attribution it is the exchange of two classes. A pair of class maps splits into a common field and a contrast, and a statistic reads only the contrast if and only if it ignores every common field; on unrestricted maps, no nonconstant such statistic is also invariant to rescaling each map on its own, so reading contrast needs an amplitude convention (Bruce, 2026f, Proposition 2.2 and Theorem 2.3). Instruments for mechanisms should be built to see the part that changes sign under the relevant involution.

Scale. Much of the controlled evidence here is below 25M parameters. The production-scale evidence is measurement on public checkpoints: captured operators at short context, sixteen pretrained language models, and the prespecified gpt-oss test, which its prespecified rule left undecided. The largest controlled test is the trained harness at 125M parameters, whose 350M leg is not yet run. 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; the compensation budget at scale is a registered test that has not run (Appendix C).

The MLP. The lattice treats the MLP as the feature writer and the main currency of compensation, and gives no theory of either role. Pure transport cannot write (Remark 1). In the tested transport family, a writing channel restores CIFAR-100 classification without restoring dictionary retrieval, 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.

Decoders. The reversible vocabulary does not apply to full causal softmax attention over more than one token, which has no positive stationary law. For decoders we use only statements proved for them: absorption at the first token and the sink-collapse law (Bruce, 2026a, Theorem 2.16), the survival reading of rollout (Bruce, 2026a, Theorem 8.5), and the two-sided law (Bruce, 2026c). 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 8). Much of what matters in decoders, including in-context learning and long-range retrieval beyond a single head, lies outside the lattice.

The cost model. The claim that softmax attention in a Pre-LN 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 and the allocator of Bruce (2026a, §§1.3 and 16.2). We did not measure the prices of the alternatives on a common hardware and optimization budget, so the fixed-point reading is an interpretation under that model, not a measured optimum.

What a successor must do. 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. Machines beyond attention are mapped against these conditions in Bruce (2026e).

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 paper.

References

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*, 2025.

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.

Neil D. B. Bruce. A spectral theory of token communication: Operators, amplification and the architecture of transformers. Zenodo, version 0.1, released with this paper, 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 paper, 2026c. URL https://doi.org/10.5281/zenodo.23541963.

Neil D. B. Bruce. Beyond transformers: Sums, maxima and the price of order. Zenodo, released with this paper, 2026e. URL https://doi.org/10.5281/zenodo.23542133.

Neil D. B. Bruce. Lost in ranking: When attribution scores hide class contrast. Manuscript, 2026f.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. arXiv:2306.15595, 2023.

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.

Aakanksha Chowdhery et al. PaLM: Scaling language modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers need registers. In *International Conference on Learning Representations*, 2024.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434, 2024.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Opher Lieber et al. Jamba: A hybrid Transformer-Mamba language model. arXiv preprint arXiv:2403.19887, 2024.

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. Curran Associates, Inc., 2025. doi: 10.52202/085713-0602.

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.

Ken M. Nakanishi. Scalable-softmax is superior for attention. arXiv preprint arXiv:2501.19399, 2025.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale Transformer training instabilities. arXiv:2309.14322, 2023.

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.

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.

Liu Yang, Kangwook Lee, Robert Nowak, and Dimitris Papailiopoulos. Looped transformers are better at learning learning algorithms. In *International Conference on Learning Representations*, 2024a.

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., 2024b.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025.

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. doi: 10.18653/v1/2025.acl-long.1126.

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.

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.

A Proofs

This appendix proves the architecture statements of Sections 2–6. Results owned by companion papers are listed in Appendix A.1; where a proof is a few lines it is repeated so that the paper can be read on its own. Every finite identity and bound below was also checked numerically on random float64 instances; the checks are falsification tests, not proofs.

A.1 Results cited from companion papers

- The operator geometry: stationary law, reversal, the split into reversible part and circulation, the lottery view of amplifiers, and the fact that row-normalizing a symmetric affinity gives a reversible walk (Bruce, 2026a, §§1.1–1.3 and 3.1–3.2).
- Collapse: short paths characterize bounded collapse and the sink-collapse law for causal operators (Bruce, 2026a, Theorems 2.8 and 2.16); retention by a bounded window, refinable laws, the mixing budget, tied loops, the Pre-LN budget and the ordering of sink treatments (Bruce, 2026a, Theorems 2.1 and 2.11, Corollary 2.12, Propositions 2.3, 2.14, 2.21 and 8.6).
- Direction and placement: the parity law (Bruce, 2026a, Theorem 3.4); the two-channel fragility identity and the anchor's cap on gain ratios (Bruce, 2026a, Theorem 4.1 and Corollary 4.2); the score decomposition into a sink and a cycle, and the cycle obstruction for a single head (Bruce, 2026a, Theorems 3.2 and 10.1).
- Feedback: the convergence certificate and its sharpness over smooth builders (Bruce, 2026a, Theorem 7.1).
- Abstention: gate times routing, the exact simultaneous budget, the sigmoid bound, tempering with separated margins, the power-mean law, the almost-sure converse and the boundary rate (Bruce, 2026c, Proposition 2, Theorem 4, Proposition 6 and Theorems 8–11).
- Allocation and compensation: the species prices (Bruce, 2026a, §1.3); the robustness premium and the corner allocator (Bruce, 2026a, §§16.1–16.2 and A.17); exact and stable compensability, subtraction across positive heads and the bounded-gain limit of compensation (Bruce, 2026a, §§5.5, 15.1, 15.5, A.6 and A.16).
- Non-expansive writes: the single-step condition for a non-expansive write, a witness that without decay the state can grow like the square root of the number of steps, the strict-contraction bound on the state, and exact linear read-back capacity (Bruce, 2026a, §§18.2–18.3 and A.19).

A.2 Structure–content separation (Section 4)

Remark 6 (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.

Proof of Theorem 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. □

Theorem 7 (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 $\|A\|_2 \le B$ and bandwidth $h \ge h_{\min} > 0$, 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 $S = D_W^{-1/2} W D_W^{-1/2}$, $S_0 = D_0^{-1/2} W^0 D_0^{-1/2}$. 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, $e^{-a'} W^0_{ij} \le W_{ij} \le e^{a'} W^0_{ij}$, the same holds for the degrees, and $\|S - S_0\|_2 \le e$; each ordered eigenvalue of S is within e of that of S_0.*

Proof. From $\mathbb{E}e^{t\|z\|^2} = (1-2t)^{-d/2}$, Chernoff's inequality gives $\|z\|^2 \le d + 2\sqrt{du} + 2u$ outside probability e^{-u}, so $\|z\| \le \sqrt{d} + \sqrt{2u}$ 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: $\|u\| \le BD$, $\|v\| \le 2\nu$ and $|\|u+v\|^2 - \|u\|^2| \le 4BD\nu + 4\nu^2$. Exponentiating gives the weight bounds, and summing them gives the degree bounds. Each normalized entry then satisfies $e^{-2a'}(S_0)_{ij} \le S_{ij} \le e^{2a'}(S_0)_{ij}$, so $|S - S_0| \le eS_0$ entrywise, and for real v, $|v^\top (S - S_0) v| \le e\, |v|^\top S_0 |v| \le e\|v\|^2$, because S_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. □

Proposition 8 (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 S_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.*

Proof. 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 $k_j = e^{-2w_j m_j^2/h^2}$, $j \in \{y, z\}$. Each has normalized eigenvalues 1 and $(1-k_j)/(1+k_j) = \tanh(w_j m_j^2/h^2)$, and the tensor product gives the four eigenpairs. Replicating the groups equally adds only zero eigenvalues. □

If the identified nuisance eigenvalue of the noisy operator is at least $t > e$, eigenvalue perturbation gives $w_z \ge (h^2/m_z^2)\,\mathrm{artanh}(t - e)$ for $0 < t - e < 1$, provided the correspondence to θ_z is established, for instance by an isolated ideal band.

Proposition 9 (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 $r \ge d_b$, and at $r = d_b$ every nonzero query q has $E^\top q \ne 0$. (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 $d_b \le r - s$.*

Proof. (i) is linearity: $\sum_j P_{ij} M k_j = M \sum_j P_{ij} k_j$, 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 $r \ge \operatorname{rank} E = d_b$; 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 $d_b \le r - s$; any isometric embedding into it attains the bound. □

Proposition 10 (Stable recovery prices approximate invisibility)**.** *Let $G \succeq 0$ be an $r \times r$ query metric with eigenvalues $\lambda_1 \le \cdots \le \lambda_r$, and let $DE = I_{d_b}$ with $\|D\|_2 \le L_D$. Then $\operatorname{tr}(E^\top G E) \ge L_D^{-2} \sum_{j=1}^{d_b} \lambda_j$, and the bound is attained.*

Proof. $\|Eb\| \ge \|b\|/L_D$ for all b. Write $E = UC$ with $U^\top U = I$, so $C^2 \succeq L_D^{-2} I$ and $\operatorname{tr}(E^\top G E) = \operatorname{tr}(C\, U^\top G U\, C) \ge L_D^{-2} \operatorname{tr}(U^\top G U)$. The last trace is smallest when U spans the d_b lowest eigenvectors of G. The choice $E = U/L_D$, $D = L_D U^\top$ attains equality. □

For a single perturbed source in fixed softmax rows the local metric is $G = \tau^{-2} \sum_i \omega_i p_{ij}(1 - p_{ij}) q_i q_i^\top$, since the Hessian of the log-partition gives $\sum_i \omega_i \mathrm{KL}(p_i \,\|\, p_i^e) = e^\top G e/2 + o(\|e\|^2)$. Width, weak query directions and decoder gain are therefore distinct resources.

Proof of Theorem 4. 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 $\sum_j p_j v_j - b_* = \sum_{j \neq *} p_j(b_j - b_*) + \sum_j p_j(v_j - b_j)$ 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 9 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 $Q\nu$.

Theorem 11 (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 $\mathrm{tr}(E^\top E) \leq P$ and $\mathrm{tr}(E^\top G E) \leq \varepsilon^2$ 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 $k = 0$.

Proof. 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 $\mathrm{tr}\, C$, 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. □

A.3 State and streaming (Section 6)

Proposition 12 (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.*

Proof. The final state is the ordered product applied to the start. Commutativity regroups it as $\prod_a A_a^{n_a} x_0$, where n_a is the number of times the symbol a occurs in the word, independent of the order, and a fixed readout preserves equality. □

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. Non-commutation can therefore be necessary in this restricted setting; it is not sufficient to solve every state-tracking task.

Proposition 13 (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).

Proof. 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 $p_{ii} \le 1$, so $1 - \alpha p_{ii} \ge 1 - \alpha = \beta$ 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 $H_{i-1} = \sum_{j<i} v_j y_j^\top$ and $H_0 = 0$, whose update is the displayed outer product; multiplying numerator and denominator by z_i gives the second form, whose denominator is $z_i - \alpha u_i^\top v_i = z_i(1 - \alpha p_{ii}) \ge \beta z_i > 0$. For a normalized kernel the weights p_{ij}, $j \le i$, sum to one, apart from any declared sink mass, so $z_i = \sum_{j \le i} u_i^\top v_j = u_i^\top s_i$ plus that mass. □

Constant memory means independent of the sequence length for fixed r and d; it does not apply to dense causal kernels, 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.

A.4 Kernels, abstention and feedback (Sections 2–5)

Proposition 14 (Logit Jacobian)**.** *For p with $p_i > 0$ and $r = \sum_i p_i \le 1$, $x^\top(\mathrm{diag}(p) - pp^\top)x \ge (1 - r)\sum_i p_i x_i^2$. Hence the matrix, which is the Hessian of $\log(e^b + \sum_j e^{s_j})$ when $p_j = e^{s_j}/(e^b + \sum_k e^{s_k})$, is positive definite if and only if $r < 1$, and it is singular with null vector $\mathbf{1}$ when $r = 1$.*

Proof. $x^\top(\mathrm{diag}(p) - pp^\top)x = \sum_i p_i x_i^2 - (\sum_i p_i x_i)^2$, and Cauchy–Schwarz with weights p_i gives $(\sum_i p_i x_i)^2 \le r \sum_i p_i x_i^2$. When $r = 1$, $\mathbf{1}$ is a null vector. □

Proposition 15 (Tied positive-semidefinite scores)**.** *If the score matrix s of a head is positive semidefinite, for instance a tied Gram form $s_{ij} = \langle W x_i, W x_j \rangle$, 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 $[0, 1]$.*

Proof. $\exp(s) = \sum_{k \ge 0} s^{\circ k}/k!$, 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 $D^{-1/2} K D^{-1/2}$, which is positive semidefinite, so its eigenvalues are real and nonnegative; they are at most one because P is stochastic. Detailed balance holds with $\pi \propto D\mathbf{1}$. □

Proposition 16 (Feedback gain with row mass)**.** *Use the maximum-row norm $\|X\| = \max_i \|x_i\|_2$ and the row-sum norm for matrices. On a closed ball $\mathcal{B}_R$ let $Q(X) \ge 0$ satisfy $\|Q(X)\|_\infty \le r_{\max}$ and $\|Q(X) - Q(Y)\|_\infty \le L_Q \|X - Y\|$. For $\alpha \ge 0$ the update $T_A(X) = \alpha Q(X) X + \beta A$ is a contraction on $\mathcal{B}_R$ when $q = \alpha(r_{\max} + R L_Q) < 1$, and for $\beta \ge 0$ with $\alpha + \beta = 1$, $\|A\| \le R$ 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 $q \le \alpha r_{\max}(1 + 2|s| R^2)$.*

Proof. $Q(X)X - Q(Y)Y = Q(X)(X - Y) + (Q(X) - Q(Y))Y$, whose norm is at most $(r_{\max} + R L_Q)\|X - Y\|$; invariance follows from $\alpha r_{\max} R + \beta R \le R$. For a sub-stochastic row p with mass r, the logit Jacobian $J_p = \mathrm{diag}(p) - pp^\top$ has $\|J_p\|_{\infty \to 1} = r - \min_{u \in \{\pm 1\}^N} (p^\top u)^2 \le r$: for a sign vector u with $m = p^\top u$, $\|J_p u\|_1 = \sum_j p_j |u_j - m| = r - m^2$, and the maximum of a norm of a linear map over the cube is attained at

a vertex. Each score changes by at most $2|s|R\|\mathrm{d}X\|$, so each row of Q changes by at most $2|s|R\,r_{\max}\|\mathrm{d}X\|$ in ℓ_1; integrating along the segment, which stays in the ball, gives L_Q. □

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.

Proposition 17 (Abstaining heads keep a stochastic core)**.** *Let $Q \geq 0$ be nonzero and irreducible, with row sums at most one and some row sum below one, with Perron root ρ and positive right and left Perron vectors h and ν. Then $0 < \rho < 1$, $P^h = \rho^{-1} D_h^{-1} Q D_h$ is row-stochastic with stationary law $\pi^h \propto \nu \circ h$, its stationary reversal is $\rho^{-1} D_\nu^{-1} Q^\top D_\nu$, and $\|Q\|_w = \rho$ in the norm $\|x\|_w = \|D_h^{-1} x\|_{\pi^h}$. If $Q_{ij} = \exp(s_{ij} - L_i)$, then $P^h = \mathrm{softmax}_{\mathrm{row}}(s + \mathbf{1}(\log h)^\top)$ and $(P^h)^\dagger = \mathrm{softmax}_{\mathrm{row}}(s^\top + \mathbf{1}(\log \nu - L)^\top)$.*

Proof. Perron–Frobenius supplies the positive vectors, and a positive root because Q is irreducible and nonzero. Multiplying $Q\mathbf{1} \leq \mathbf{1}$, strict in some row, by $\nu^\top$ gives $\rho\,\nu^\top \mathbf{1} < \nu^\top \mathbf{1}$, so $\rho < 1$. $(Qh)_i = \rho h_i$ gives unit row sums of P^h. Stationarity: $\sum_i \nu_i h_i \rho^{-1} h_i^{-1} Q_{ij} h_j = \rho^{-1} h_j (\nu^\top Q)_j = \nu_j h_j$. The reversal entries are $\pi_j^h P_{ji}^h / \pi_i^h = \nu_j Q_{ji} / (\rho \nu_i)$. Since $Q = \rho D_h P^h D_h^{-1}$, $\|Qx\|_w = \rho \|P^h D_h^{-1} x\|_{\pi^h} \leq \rho \|D_h^{-1} x\|_{\pi^h} = \rho \|x\|_w$, because a stochastic matrix is a contraction in the geometry of its stationary law, with equality at $x = h$. For the softmax forms substitute $Q_{ij} = e^{s_{ij} - L_i}$ and use $\rho h_i e^{L_i} = \sum_j e^{s_{ij}} h_j$ and $\sum_j \nu_j e^{s_{ji} - L_j} = \rho \nu_i$. □

Strict killing in some row is not enough without irreducibility, since a disconnected stochastic block keeps eigenvalue one; survival at most $r < 1$ in every row gives $\|Q^t\|_\infty \leq r^t$ directly. A causal head is reducible, and its conditioned behaviour needs an accessibility check. For an absorbing chain whose transient block T reaches the target from every state, taking the longest of the finitely many shortest absorbing paths and the smallest of their positive probabilities gives $\|T^m\|_\infty < 1$ for some finite m, so all mass is absorbed. Conditioned on survival, however, a causal chain concentrates on the transient token with the largest self-attention only if that token is reachable from its start: $Q = \mathrm{diag}(0.9, 0.5)$ started in state 2 stays there whenever it survives.

B Evidence standard and validity

Tiers. Measurements carry the monograph's four tiers (Bruce, 2026a, Appendix B). **ID** is a machine-precision check of an identity; it certifies algebra and code, not a claim about models. **SYN** is a law checked on synthetic constructions. **SYN-R** is a law checked on real operators or spectra along an artificial perturbation. **REAL** is a measurement of trained models or real data, including models trained on synthetic tasks. Claims about practice carry two further labels. A **retrodiction** reads an adopted change as a lattice move after the fact (Table 5); it names a lattice quantity but is never evidence for the lattice. A **prediction** is stated before any test of it (Section 9); only tests made after it is stated can count as evidence for it.

What "established" means. A statement is established here when it is proved, with its hypotheses stated, or when an experiment preregistered with a falsifier returned the registered outcome on its frozen criteria and its validity checks held, at the tier given. A pass of an implementation gate is not scientific support: a cell can pass every correctness check while the effect it was built to measure is absent (Bruce, 2026a, §R.2). Where a later audit corrected a record or an assessment, the frozen verdict is retained as provenance, with the correction stated beside it.

Dispositions. Verdicts use the program's vocabulary: pass, fail, mixed, inconclusive, void, withdrawn and rescoped, with data-absent for a leg whose inputs never arrived. A void or data-absent leg is never read as a negative. For an inconclusive result, the account states what the measurements show and what remains unresolved, including any trend they support. Tolerances and gates are fixed before the run; corrections to defective gates are recorded as amendments, with the original criteria and verdicts retained; smoke runs are excluded; and where a saved assessment is stale, the quoted numbers are recomputed from the per-cell records.

How quoted records are scoped. Every number quoted from a program record is checked against that record and quoted at the scope the record supports. The looped-model result is quoted per arm and per seed, not pooled across the moving-average arms (Table 1). The value-tying tax is quoted on one slice of depth and width; in four-layer models it persists at MLP ratio 4. The 2,700 causal operators are captured heads and head-averaged layers at short context, not production measurements. The collapse horizon of graph convolutions is an operator quantity, not a trained-model accuracy. The one test of a coupled logit scale is not cited for gain control: its toy loop's flags tracked the anchor's relaxation, and its byte model's detector took normal attention sharpening for instability. The 22M-parameter minimal-block comparison is not cited as dominance until its tuned replication returns.

Provenance. Table 6 lists every experiment held in this package that the paper quotes or relies on, with its evidence tier, where it is quoted, and the folder of this package's `data/` directory that holds its records and their checksums. Measurements cited to the companion papers are held in their packages, and the two negatives that Section 8 says rest on weaker records are not in this package.

Table 6: Experiments quoted in the paper, with their evidence tiers and the folders that hold their records.

Experiment and what is quoted	Tier	Where	Data folder
Identity path removed from the minimal block: exact chance at depth ≥ 12	REAL	§3	`data/N25`
Attention sublayer's identity path removed in a 22M-parameter block: exact chance, constant output	REAL	§3	`data/E2.1`
Collapse horizon of graph-convolution operators (shortest: 87, a Cora subgraph)	SYN-R	§3	`data/E3.2`
Trained graph convolutional networks collapse at depth 8–16	REAL	§3	`data/E05`
Post-LN trainability cliff at depth ≥ 18 (CIFAR-10, bimodal)	REAL	§3	`data/E04`
Looped models with and without input re-injection, per seed (Table 1)	REAL	§3, 6	`data/SX43`
Tied value map fails at head width $\leq d_b$; recovery above it	REAL	§3, 4, 7	`data/N12`, `data/N12b`
Value-tying tax $+1.70/+0.04/-0.72$ pp at MLP ratio 0/1/4 (CIFAR-100, depth 24, width 192)	REAL	§3, 7	`data/E04b`
Synthetic checks behind the cited feedback results: past the certified margin, three fixed points, the outer two attracting, with hysteresis; normalization removes an adaptive bandwidth's runaway	SYN	§3	`data/N31`, `data/N33`
Sigmoid attention: row mass 0.31–0.34; sink incidence	REAL	§3	`data/N19`
Output gating halves content-sink mass in a small hybrid model (causal-detector caveat)	REAL	§3	`data/R7c`
Census of 2,700 captured causal operators; median 0.416	REAL	§3	`data/RIII1`
Retrieval accuracy onset $\exp(2.14 \cdot \text{gap})$, 15 of 15 cells	REAL	§3	`data/R5d`
$\log N$ re-score on the resolution axis (trained from scratch; upsampled test images)	REAL	§3	`data/V1`
$\log N$ re-score of trained ViT-S checkpoints on images upsampled to $16\times$ the training tokens; plain softmax gains at least as much as the tempered sink	REAL	§3, 8	`data/SX102`
The same checkpoints re-evaluated: the re-score gains $+4.5$ to $+4.9$ pp on resized original images (most upscaled at $16\times$); the rows dilute, and tempering moves them away from their 224 px pattern; the learned sink's log-odds fall at about -0.96 per e-fold, and abstention holds level with an output gate (reported, not scored)	REAL	§3, 8	`data/SX115`
Tied (Gram) scores copy less surely under a causal mask; parity on bidirectional lookup	REAL	§3	`data/N13`
Pure transport cannot write (dictionary retrieval, CIFAR-100); a write channel restores classification only	REAL	§3, 8	`data/ATS07`, `data/ATS12`
Engaged tied excess risk tracks the dynamic range	SYN	§4	`data/N11`
Range-contained head at risk 0.311–0.312; noise floor 0.079	SYN	§4	`data/E2.2`
Compact or independent value channel: cache saved, no gain in task performance (trained models) or decoding speed (a timing benchmark on newly initialized heads and random inputs)	SYN, REAL	§5, 8	`data/ATS08`, `data/ATS14`, `data/ATS15`
Delta-rule write certificate checked to machine precision: its predicate matched non-expansiveness in all 200,000 random draws	ID	§2, 6	`data/E4.2`
Certified fast-weight memories: divergence against the certificate, and a synthetic memory swept across its boundary	SYN, REAL	§6	`data/AF01`
Decay gate removed and state register reserved in a trained hybrid: state growth, perplexity, recall (local records, reconciled)	REAL	§6	`data/N36`
Gated delta-rule recurrence under a matched-data protocol: recall of stored pairs	REAL	§6, 8	`data/ATS04`
Certified non-expansive flow block: degradation in bits per byte at eight times the training context, in-sample (Bruce, 2026a, §18.6)	REAL	§6, 8	`data/FB7`
Rank clause exact on its construction; evaded by a content map trained with its readout	REAL	§7	`data/N24b3`

Experiment and what is quoted	Tier	Where	Data folder
Subtraction as a head-capacity resource	REAL	§7	`data/BD30`
Twelve explicit-knob interventions: mechanisms held in several rows, not in all; capability null or negative in every row that could be scored	SYN, REAL	§7	`data/AF01` to `data/AF12`
Corner allocations; capped greedy ordering fails	SYN	§7	`data/N35`
Deployed-design record (descriptive)	REAL	§7	`data/E5.4`
Detailed balance in trained minimal encoders; 20/20 parity arm	REAL	§7	`data/N24`, `data/N24b2`
Propagate-once aggregation	REAL	§7	`data/N27`
Conditioning-gated amplification, three field tests in graph label propagation (26 synthetic and 4 real graphs) and a synthetic node-classification system: a regressed dial loses, a measured dial ties within tolerance, the certified allocator wins	SYN, REAL	§7, 8	`data/N28c`, `data/N28d`, `data/N38`
Signed gain shaping for relevant modes at low or negative eigenvalues	REAL	§7	`data/N29`
Sink length law in a production model (gpt-oss-20b): inconclusive, leaning toward compensation through position without settling it	REAL	§8, 9, App. C	`data/SX01`
Trained harness of 125M-parameter language models and its vision subset: abstention rules, sinks, massive activations and length (Figure 3)	REAL	§5, 7, 8, 9, App. C	`data/SX02`
Reversal objects leak future tokens in a decoder: a frozen GPT-2 with flow-and-resolvent layers reaches validation perplexity 18.09, against 23.43 with the reversible part and circulation causally re-masked and 23.50 for the resolvent alone (one seed)	REAL	§8	`data/decoder_leak`
Recall test of a minimal recurrent arm: bimodal across seeds with its curriculum on, at a learning rate not retuned for it	REAL	§8	`data/FB8`
Organized blocks' construction identities checked to machine precision	ID	§8	`data/FB0`
Organized blocks roughly match a parameter-matched transformer on three small causal tasks	REAL	§8	`data/FB6`
Engineered null registers: mean content-sink mass cut by 14% (causal-detector caveat), no detectable advantage over a plain register token	REAL	§8	`data/FB10`
Cache eviction by absorbed mass loses to heuristic eviction, judged in a linear model built from captured attention	SYN-R	§8	`data/AX31b`
Free signed normalization: no win on heterophilous graphs, harm on homophilous ones	REAL	§8	`data/BD28`

C Registered tests

Each test below has a protocol, a prediction and a decision rule fixed before it runs, frozen in the repository rather than in an external public registry. Two have returned: the test of the sink length law in a production model, which leaned toward compensation through position without settling it, and the abstention harness, at 125M and in its vision subset. The others have not run. Where the text depends on one of them it says so, and Table 7 gives what each measures and what each outcome would change in this paper.

Table 7: Registered tests: what each measures, its status, and what each outcome would change in this paper.

Test	What it measures	Status	What each outcome would change
Sink length law in a production model. Do learned sinks stop abstaining as context grows?	In gpt-oss-20b, the slope of the sink mass's log-odds against the logarithm of the query position, in dense layers and in 128-token banded layers, which serve as a built-in control; a token-shuffled copy of each document controls the background.	*Returned; undecided by its prespecified rule, the slopes leaning descriptively toward compensation.* The books shuffle control's gap was 0.672 nats, below the required 1 nat (the natural books text scored 7.06 nats per token), and the slopes (-0.20 in dense layers, -0.22 while the band fills, -0.02 beyond it) matched neither prespecified branch. The prespecified protocol, its gates and its deviations are in Bruce (2026c, Appendix C).	A confirmation would have given the two-sided law a test in a production model, and compensation would have moved the claim to the retrieval side. As returned, the claims about trained models rest on the controlled pilot and the trained harness (§8; prediction 1).

Test	What it measures	Status	What each outcome would change
Abstention harness. Which abstention rules keep both abstention and retrieval as context grows, and what does dropping unit row sums cost at depth?	Decoders trained at 1,024 tokens with each of 11 attention rules and evaluated to 8× that length for abstention, retrieval, first-token sinks, massive activations and perplexity; a subset of the rules in vision transformers; and a depth ladder to 48 layers for the gain-control cost.	*Returned at 125M* (3 seeds per rule) and for the vision subset. At 125M, tempering on top of YaRN left the perplexity ratio at 8× +0.27 and +0.18 above softmax's and did not hold abstention within the registered bound, and every rule with a null removed the first-token sink and cut massive activations by 67% to 82%. In the vision subset the one registered prediction failed: content-sink mass fell in every seed, the massive-activation ratio in only two of three. The 350M models and the depth ladder have not run.	At 125M the lead prediction failed: tempering on top of YaRN is not a fix in these language models, though on interpolated positions it helped, while every null removed the first-token sink and most massive activations there, as registered for two rules (§8). At 350M: whether both results hold with scale. On the depth ladder: whether dropping unit row sums obliges norm control at depth (rows A1 and C3).
Anchor horizon at scale. Does a measured anchor strength predict where trained deep and looped transformers lose content?	Looped blocks with input re-injection at several weights, and deep untied models with scaled residual branches: the per-visit mixing share, the symmetric fraction and the horizon, with the anchor also varied at evaluation. Blocks with the identity path removed, given more capacity or a re-supplied anchor: whether capacity can stand in for the anchor.	Not run; small looped models are in Table 1.	Confirmation: the anchor's price law governs recursion depth in trained transformers (prediction 9; row C1). Refutation: the law is scoped to operators.
Coupled logit scale. Does coupling the logit scale to weight decay predict and prevent logit instabilities?	In a 125M language model across a learning-rate sweep, a spectral cap on the query–key maps tied to weight decay against QK-norm, soft-capping, z-loss and σReparam, with the loop gain measured directly.	Not run.	Confirmation: the certificate predicts where the standard fixes are needed, and the coupling matches the best of them at parity (prediction 7; §8.5). Refutation: the certificate explains the fixes but does not predict where instability begins.
Compensation at scale. Does capacity still absorb the value-tying tax at scale?	The tax of a tied value map against MLP ratio, width and depth, in vision transformers on ImageNet-100 and in a 125M language model.	Not run.	Confirmation fits a compensation budget at scale and supports the flat neighbourhood of §7; a tax that stays positive at MLP ratio 4 makes the alignment protection necessary at scale (row C2).
Head-count sweep. Do gains from signed attention shrink as heads multiply?	Softmax, differential and signed-gate attention with 2 to 16 heads at fixed width, in a 125M language model and in ViT-S.	Not run.	Confirmation makes subtraction a head-capacity resource that signed attention supplies only where heads are scarce (prediction 4; rows A2 and C5). Refutation: signed mixing adds value beyond subtraction across heads.
Circulation gates. Can a model use an explicit per-head direction knob, and is its learned setting predictable?	Learned gates on each head's circulation in pretrained vision transformers, against temperature and attenuation knobs of equal capacity, with the sign of each learned change compared with the first-order predictor.	Not run.	Confirmation makes direction a controllable knob, the sign of whose learned change the first-order predictor anticipates (row A3; §8). Refutation: the knob is used no better than generic controls.

Test	What it measures	Status	What each outcome would change
Reversible composition. Do two reversible mixers learn to compose a directed block?	Pairs of reversible mixers, with and without a cyclic coordinate shared by both layers, trained on successor routing, with a single mixer and commuting pairs as controls.	Not run.	Confirmation: given that coordinate, depth stands in for circulation, which compensates the arrow (row A3). Otherwise direction in this setting comes from a circulating head or from computation with an MLP (§3).
Diffusion language models. Do masked diffusion language models carry more circulation than masked-LM encoders of similar size?	A per-head circulation census of a masked diffusion language model and masked-LM encoders, with a reversal probe on next-line and previous-line recall.	Not run.	Agreement: a model that generates without a mask builds its arrow as circulation (prediction 5; row A3). Disagreement: the prediction is not supported at that scale.
Minimal block against tuned baselines. Does the minimal anchored block's margin survive tuned baselines?	The 22M-parameter record completed, the block re-implemented cleanly and compared with DeiT-recipe ViT-S and ResNet-50 baselines at matched parameters on ImageNet-100, with one-factor ablations.	Not run.	A mean margin of at least 2 pp, positive in every paired seed, gives the blueprint an at-scale data point and names the factor that carries it; parity is stated as parity; a loss leaves the blueprint on its tiny-scale comparison (§7.1).
Hybrid ratio sweep. Is a protection-complete hybrid a resource-axis win at quality parity?	The fraction of attention layers, from none to all, in 125M and 350M language models whose recurrent layers carry a certified contracting write, against a qualified attention baseline: quality, degradation at 8× the training length, recall and state size, with an uncertified hybrid to isolate the certificate.	Not run.	Parity at a quarter attention, with less degradation and less state, makes the hybrid a resource-axis win and measures the certificate's share; otherwise the trade-off frontier is reported. Either way it shows whether the best fraction sits at an endpoint or a cap, as concave prices predict (row B2; §8).