Attention Sinks at Long Context: One Log-N Law for Retrieval and Abstention

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

Abstract

Softmax attention cannot abstain on its own: its weights sum to one, so a head with nothing relevant to read must still attend somewhere. Trained models improvise such a place, the attention sink. Practice now builds one in: a null, an option that absorbs attention and delivers nothing, such as a learned sink logit, a register with zero value or an output gate. We show that these nulls are one mechanism, a gate on an ordinary softmax, and ask what happens to them as the context grows. The answer is one dilution law for two abilities. To retrieve, a matching key must outweigh the N other keys; to abstain when nothing matches, the null must. With bounded scores both leads must grow like $\log N$, and because a query with a match and one without share the same null, no rule that sums over the keys keeps both at long context: a fixed null loses abstention, and a null that grows with the length loses retrieval. Tempering, which scales every score, the null included, by $\log N / \log N_{\text{train}}$, provably keeps both under margin conditions, and a head that decides by its largest score escapes the law. Trained heads in a controlled pilot follow the law closely, and tempering keeps both sides there, though not in every harder setting. In 125M-parameter language models, every rule with a null removed the first-token sink and most massive activations; past the training length, keeping positions in range mattered more than the null rule, and adding tempering to YaRN, which already includes a temperature, hurt. In small vision transformers trained at one resolution, tempering at evaluation raised accuracy at higher resolutions. A prespecified test on gpt-oss was inconclusive.

1 Introduction

A softmax attention head spends a fixed unit of mass. Its rows sum to one, and the only change of the logits that would mean "attend less", a uniform shift of the whole row, is the one direction in the kernel of the softmax Jacobian: it changes nothing (Section 3). A head with nothing to retrieve must still put its mass somewhere. Trained language models manufacture places to put it: sink tokens, often the first, that absorb attention while contributing little (Xiao et al., 2024; Sun et al., 2024; Gu et al., 2025; Barbero et al., 2025). Practice has since built the escape in directly, with register tokens (Darcet et al., 2024), off-by-one softmax (Miller, 2023), learned per-head sink logits (OpenAI, 2025; Wang, 2026), headwise output gates (Qiu et al., 2025), sigmoid attention with a $-\log N$ bias (Ramapuram et al., 2025), and rectified or stick-breaking alternatives to softmax (Zuhri et al., 2025; Tan et al., 2025). Of these, only sigmoid attention's bias depends on the context length.

This paper asks what happens to these mechanisms as the context grows, and its answer is one dilution law for two capabilities: retrieval, putting the weight on the one key that matches, and abstention, putting it on the null, the option that absorbs attention and delivers nothing, when no key matches. Its setup pairs two matched queries that see the same null and the same N background keys, the keys that do not match. The present query also has a matching key; the absent query has none. In each, a single logit must beat the N background logits: the matching key's logit to retrieve, the null's to abstain. As N grows, that logit's share is diluted unless its lead over the background grows, so with a bounded logit scale both leads must grow like $\log N$. They compete, because the two queries share one null: a null strong enough to win when no key matches also takes weight from the match when one does. Theorem 4 makes this exact: if every content score,

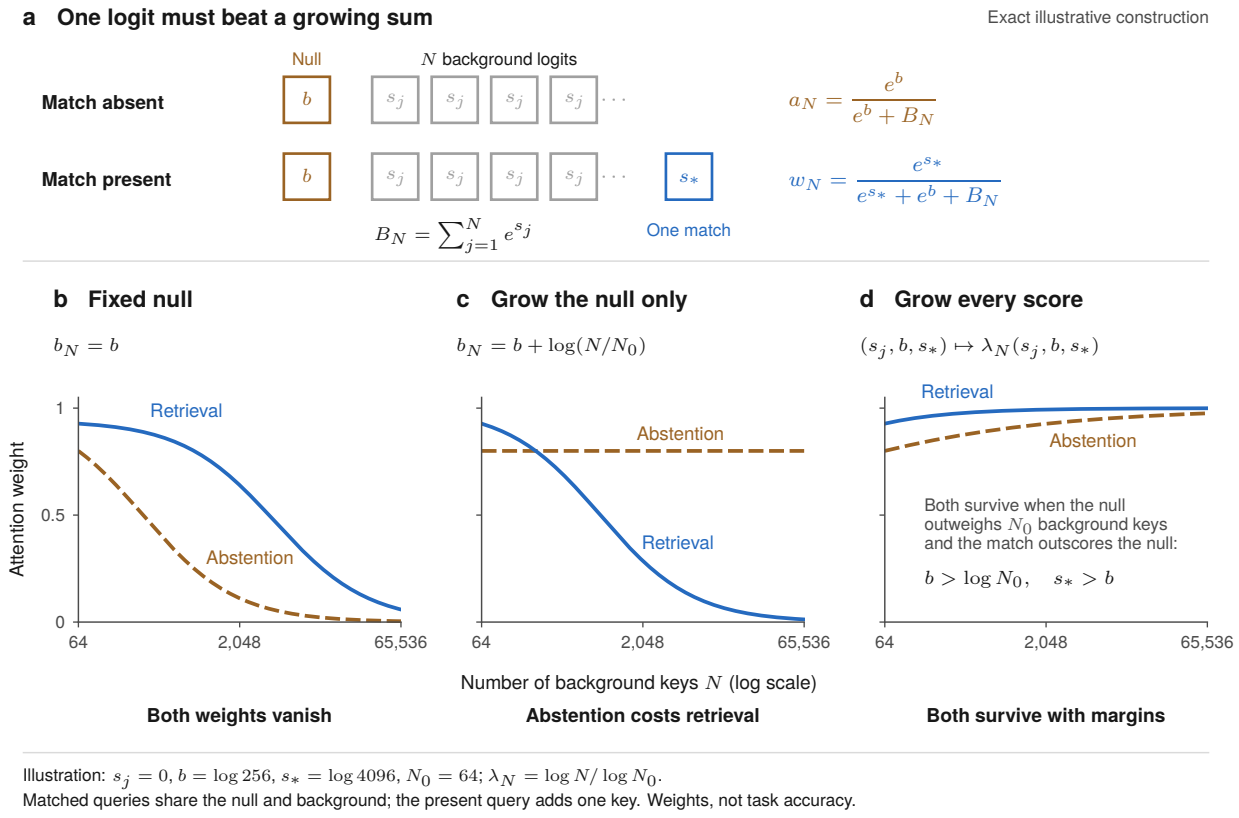


Illustration: $s_j = 0$, $b = \log 256$, $s_* = \log 4096$, $N_0 = 64$; $\lambda_N = \log N / \log N_0$.
Matched queries share the null and background; the present query adds one key. Weights, not task accuracy.

Figure 1: Growing the null alone trades retrieval for abstention; growing every score can preserve both when the null and the match keep separated margins over the background. **(a)** Matched absent and present queries share N background keys with scores s_j and a null score b; the present query adds one match with score s_*. Squares denote scores, not weights. The background sum is $B_N = \sum_j e^{s_j}$. Brown dashed curves show the absent query's null weight a_N; blue solid curves show the present query's match weight w_N. **(b)** Fixed scores eventually lose both weights. **(c)** Adding $\log(N/N_0)$ only to the null holds abstention constant while retrieval vanishes. **(d)** Multiplying every score, null included, by $\lambda_N = \log N / \log N_0$ preserves both in this exact illustrative construction: $s_j = 0$, $b = \log 256$, $s_* = \log 4096$, $N_0 = 64$. Its sufficient-margin factors are $N_0 e^{-b} = 1/4$, $N_0 e^{-s_*} = 1/64$ and $e^{b-s_*} = 1/16$, all below one. These are attention weights, not task accuracies; no measurements are drawn. Section 6 tests the three regimes in trained heads, where tempering's success is measured rather than guaranteed. The law covers rules that decide by summing over the keys; a gate that sees whether the match is present, or a head that reads only the largest score, lies outside it (Sections 5.1 and 5.5).

the score of a key rather than of the null, lies in $[-\tau, \tau]$, no choice of null (fixed, learned, or calibrated to the length) keeps abstention (the absent query's weight on the null) above a_* and the match weight (the present query's weight on its match) above w_* beyond $N_* = e^{2\tau}(1-a_*)(1-w_*)/w_*$, and the bound is sharp. A fixed sink eventually loses both abstention and retrieval. A sink calibrated to the length, or sigmoid attention with a $-\log N$ bias, keeps abstention and loses single-key retrieval.

The way out is to grow the whole logit budget, the range $[-\tau, \tau]$ within which the content scores lie: a logit budget that grows like $\log N$ is necessary whenever the two queries share one null and each decision sums over the keys (Theorem 4). *Tempering* multiplies every logit, the null included, by $\lambda = \log N / \log N_0$, where N_0 is the training length. It keeps both sides when the null and the match hold separated margins over the background (Theorem 8). Before this scaling, each must clear every background score, at every length, by at least $\log N_0$, and the match must outscore the null. Figure 1 compares fixed scores, a null raised by $\log(N/N_0)$ alone, and tempering on an exact construction.

The law is sharp, and it says exactly where tempering fails. With one common multiplier on every score, abstention obeys an exact power-mean law: it depends on the background only through one power mean of the exponentiated background scores, which lies between their ordinary mean and their maximum (Theorem 9). For independent, identically distributed background scores, abstention converges almost surely to one, to a boundary value, or to zero as the null sits above, at, or below a threshold: the essential supremum of the background score plus $\log N_0$ (Theorem 10). The essential supremum is the smallest upper bound on the background score that holds with probability one. If the null only matches a uniform background score, abstention is $1/(N+1)$ at every temperature.

A controlled pilot trains single attention heads at two logit budgets and two training lengths, to retrieve a matching key when there is one and to abstain when there is none, and evaluates them to 65,536 keys (Section 6). At the smaller logit budget and training length, a fixed learned sink's abstention falls from 0.874 to 0.007 along the prediction of the mean-field law (within 0.002 per seed), which replaces the background sum by its expectation, and the length-calibrated rules, a sink raised by $\log N$ and sigmoid attention with a $-\log N$ bias, lose retrieval from $N = 2{,}048$. A larger logit budget or training length moves these failures later or past the tested range. The tempered rule keeps both over the tested range, although its trained margins over the background do not meet the conditions of Theorem 8. We call a protocol registered, or prespecified, when its analysis was frozen, with hashes, before its run; none of the protocols here was publicly deposited. A registered extension to seven further settings (clustered or heavy-tailed keys, near matches at three cosines, and continuous values read by one head or by four 32-dimensional heads) keeps this pattern for rules that sum over the keys, with two qualifications. With heavy-tailed keys every such rule loses a side, tempering included. With clustered keys at the smaller logit budget, tempering's abstention accuracy falls below 0.9 inside the tested range. Beyond the registered checks, at that budget tempering also loses abstention with four 32-dimensional heads, as its own law predicts for their wider background (in 32 dimensions, random cosines spread twice as widely as in 128), and with near matches the tempered head does not abstain even at the training length.

Heads that decide by reading the largest score, instead of a sum over the keys, need no $\log N$ schedule against a background bounded below their threshold (Section 5). In the extension, a head that reads the largest score for both retrieval and abstention keeps both sides except with heavy-tailed keys. A head that reads it only to abstain, and routes over every key for retrieval, also loses retrieval at the smaller logit budget with a near match at cosine 0.60.

Contributions. Each contribution carries two labels: how it is shown, proved or measured, and how far it reaches: exactly, under premises that must be checked head by head, in an idealized setting, or in trained networks.

- **The constraint and the class** (Sections 3 and 4). *Proved; exact.* A null logit turns the softmax Jacobian from singular to positive definite, and every sink-logit or headwise-gated head is a diagonal gate on an ordinary softmax. Sigmoid attention lies outside this class. An abstaining head is a killed Markov chain: each row loses the mass assigned to abstention. When it is irreducible, as every unmasked head is, the

chain conditioned on never abstaining, and that chain's time reversal, are again attention calls. A causal mask already supplies an absorbing token.

- **The two-sided law** (Section 5). *Proved. Under premises that must be checked head by head: content scores in $[-\tau, \tau]$ and a null shared by the matched queries (the exact condition and the bounds for each rule), separated margins over the background (tempering), and one multiplier on every score (the power-mean law). In an idealized setting: the converse and the mean-field law, which assume independent, identically distributed background scores.* The exact condition for one null to keep both sides and its sharp impossibility bound; the bounds for fixed sinks, calibrated sinks and sigmoid weights; the tempering theorem with its separated-margin conditions and a counterexample; the exact power-mean law and its almost-sure converse; and a mean-field law with finite-N corrections.
- **A controlled pilot** (Section 6). *Measured, in an idealized setting: trained heads on constructed keys, one head per rule except in one setting with four.* The pilot and its registered extension to seven further settings.
- **Trained models** (Sections 7 and 8). *Measured in trained networks: small models trained for the purpose, up to 125M parameters, and observational measurements of pretrained ones.* Sink measurements in small trained models, a census of seven pretrained vision transformers, and a prespecified trained harness of vision transformers and 125M-parameter language models. In the harness, tempering every logit by $\log N / \log N_{\text{train}}$ on top of YaRN, which already carries a log-length temperature, did not keep the null's attention weight (at this scale no rule abstained in behaviour) and raised long-context perplexity, while every rule that gives attention a null removed the first-token sink and most massive activations. Tempering at evaluation raised the vision transformers' accuracy at up to 16 times the training token count, and in the language models interpolating the rotary positions alone carried most of YaRN's rescue; Section 11 draws the recommendations for practice.
- **A production model** (Section 9). *Measured in a trained production network.* A prespecified test asks whether the odds of gpt-oss-20b's sinks against the background fall as the visible context grows, as the fixed-sink law predicts, with its banded layers, whose background stops growing at 128 keys, as a built-in control. It was undecided by its prespecified rule, because its control comparing natural and shuffled books failed and its sink log-odds slopes matched neither prespecified branch; descriptively, the slopes lean toward compensation through position, a network's lowering of its background logits as the context grows.

What is not new. Sinks are not new, and neither is the retrieval half of the law. That a winning logit needs a lead over the background growing like $\log N$ is known (Nakanishi, 2025; Veličković et al., 2025), and practice already scales logits with length (Chiang & Cholak, 2022; Peng et al., 2024; Bai et al., 2023; Nakanishi, 2025; Anson et al., 2025); recent theory locates the critical scaling in the upper tail of the score distribution (Hayase & Karakida, 2026; Bruno et al., 2026). None of this work treats abstention, and none of it tempers the null. Two recent designs do let the null follow the length: one tempers every score, its null included, by a learned factor and raises the null like the background's extreme value (Akbar, 2026), and the other trains a sink calibrated to the corpus size (Gollapudi et al., 2026). What is new here is the abstention half in general form, the two-sided failure, the conditions under which tempering the whole logit budget keeps both sides, and the tests. We propose no new attention mechanism and claim no state-of-the-art result. The law concerns rules whose decisions sum over the keys at a fixed logit scale; a head that reads a maximum instead escapes it (Bruce, 2026e). The monograph (Bruce, 2026a), the reference text for this paper and its companions, develops the operator theory beneath this paper and holds the companion results it cites.

2 Related work

Sinks and why they form. Language models place large attention on their first tokens, and keeping those tokens' keys and values rescues windowed attention (Xiao et al., 2024). Vision transformers develop high-norm tokens in uninformative background patches, which added register tokens absorb (Darcet et al., 2024). A few massive activations act as fixed biases and concentrate attention on their tokens (Sun et al., 2024). Sinks emerge with optimization and data and depend on the softmax normalization (Gu et al., 2025), and they may help a deep model avoid over-mixing its tokens (Barbero et al., 2025). Our starting point is

the constraint that makes a place for surplus mass necessary (Section 3). The production census in Section 7 adds that in pretrained vision transformers the in-sequence sinks it detects are not empty, though selected for small value norms: they carry content of their own.

Mechanisms that let a head abstain. Off-by-one softmax adds a null key of logit zero (Miller, 2023); gpt-oss learns a sink logit per head (OpenAI, 2025); and, concurrently with this work, Wang (2026) learns a per-head sink logit, a phantom key with zero value, for abstention and adds value gates for noise filtering. In models of 10M to 350M parameters, trained and evaluated at 512 and 1,024 tokens, the sink logit removes the first-position sink. The benefit of gating comes almost entirely from abstention at 10M but is shared equally with filtering by 350M. The study also reports the sink mass of gpt-oss-20b, averaged over positions. Its Lemma 3.1 decomposes a gate on the values into a renormalized routing times a row attenuation, the value-gate counterpart of our factorization of sink-logit heads (Proposition 2). The study does not vary the context length. Bondarenko et al. (2023) traced outliers to heads that try to make no update, and removed them with a clipped softmax and with a gate on each head's output; output gates driven by the query also remove sinks and massive activations in large language models (Qiu et al., 2025). Rectified softmax drops the sum-to-one constraint (Zuhri et al., 2025), and stick-breaking attention leaves the unallocated remainder of each row unused (Tan et al., 2025). Sigmoid attention replaces the softmax by elementwise sigmoids with a $-\log N$ bias (Ramapuram et al., 2025). In Section 4 the sink-logit and off-by-one heads, heads with value-null registers (registers whose value is zero), and headwise output gates become one class of diagonal gates on an ordinary softmax; sigmoid attention falls outside it. Among these fixed rules only sigmoid attention's bias depends on the length, and it is chosen to hold the row mass fixed, which the law shows costs single-key retrieval. Two recent designs let a softmax null follow the length. ATMA's polar attention multiplies every score, its null included, by a learned factor $1 + a \log n$. It raises the null's logit like $\sqrt{\log n}$, the extreme-value scale of unrelated scores; both are learned per head. For sub-Gaussian scores it bounds the noise that rises above the null (Akbar, 2026). BlockSearch trains a sink per layer whose threshold grows as $\log(N/N_0)$ with the corpus size and switches it off at evaluation. Switching it back on left retrieval on MS MARCO essentially unchanged; scaling every score by $s \log N$ instead closed most of the gap to a dense retriever (Gollapudi et al., 2026). In the law's terms, BlockSearch's sink is a null calibrated to the length, which cannot restore single-key retrieval (Section 5), and ATMA pairs tempering with a null that tracks the background's extreme value.

Scaling logits with length. Multiplying the logits by $\log n$ keeps attention sharp as the input grows (Chiang & Cholak, 2022), and a factor $\log n / \log n_{\text{train}}$ keeps the attention entropy roughly invariant (Su, 2021). Practice applies such factors beyond the training length (Bai et al., 2023; Meta AI, 2025), folds a temperature into position interpolation (Peng et al., 2024), or learns the scale (Nakanishi, 2025); scale-invariant attention transforms the logits by position so that total attention and its sparsity stay fixed (Anson et al., 2025). Recent theory places the critical scaling in the upper tail of the score distribution, through a gap-counting function (Hayase & Karakida, 2026) or the local behaviour of the distance to the query (Bruno et al., 2026), rather than in a universal $\log n$; for Gaussian logits it is of order $\sqrt{\log n}$ (Giorlandino & Goldt, 2026; Chen et al., 2026, Appendix D). The placement is classical for sums of random exponentials, where it is the freezing transition of the random energy model, the point past which the sum is carried by its largest terms (Derrida, 1981; Janßen, 2010). The same logit budget appears on the resolution axis of vision models, as log-token scaling, test-time temperatures and proportional attention (Wang et al., 2024; Leroy et al., 2024; Gao et al., 2024). All of this work concerns retrieval, the sharpness of attention on relevant keys. Outside softmax nulls, the clipped softmax's floor, above, has been scaled with length (Bondarenko et al., 2023), a thresholded attention tempers its threshold with length so that a row may keep nothing (Huang et al., 2026), and sinks have been measured against training and test length (Barbero et al., 2025; Gu et al., 2025; Qiu et al., 2025). We are not aware of work that treats a softmax null's abstention and retrieval as two sides of one logit budget; the two length-dependent nulls above each impose a form of growth, and ATMA's analysis bounds the noise above its null and the total noise-to-null odds, but not the match. Theorems 10 and 11 are the abstention counterparts of the upper-tail results. Under tempering, the phase boundary for the null logit sits at the background's essential supremum plus $\log N_0$, and the upper tail sets the rate of convergence at the boundary.

Caches. Sinks also decide which tokens a bounded cache must keep: windowed attention keeps them (Xiao et al., 2024), heavy-hitter eviction keeps the tokens that receive the most attention (Zhang et al., 2023), and

TOVA drops the tokens that receive the least (Oren et al., 2024). We touch eviction only to report one negative result. A prespecified test used a policy built from absorbed mass, which measures how much of each cached token's content had drained into the sink through the layers. The policy lost to all three, judged in a linear model built from captured attention (Sections 7 and 11).

3 The constraint: sum-to-one leaves no direction for attending less

Fix one query and write its N logits as $s \in \mathbb{R}^N$. A softmax head assigns $p = \mathrm{softmax}(s)$, $p_j = e^{s_j} / \sum_k e^{s_k}$. Its Jacobian is

$$J = \frac{\partial p}{\partial s} = \mathrm{diag}(p) - pp^\top, \tag{1}$$

which is positive semidefinite, and for positive weights its kernel is exactly the span of $\mathbf{1}$. The kernel is the uniform shift $s \mapsto s - c\mathbf{1}$, the one change of the logits that could mean "attend less", and it changes nothing. A softmax head cannot lower its total attention, and gradient descent cannot teach it to: the direction carries no gradient.

A null logit b changes this. The head becomes

$$p_j = \frac{e^{s_j}}{e^b + \sum_k e^{s_k}}, \qquad r = \sum_j p_j < 1, \tag{2}$$

where the missing mass $1 - r$ is the head's abstention. Off-by-one softmax is $b = 0$ (Miller, 2023); a learned sink is a trained b per head (OpenAI, 2025). Several value-null registers, whose values are zero, enter through their summed exponential weight. The Jacobian keeps the form (1), now with $\mathbf{1}^\top p = r$. Write $\mathrm{LSE}(s) = \log \sum_k e^{s_k}$ for the log-sum-exp and $\sigma(z) = 1/(1 + e^{-z})$ for the logistic function.

Proposition 1 (A null makes the Jacobian definite)**.** *For the head* (2) *and every* $x \in \mathbb{R}^N$*,*

$$x^\top J x \;\geq\; (1 - r) \sum_j p_j x_j^2,$$

so J *is positive definite with* $\lambda_{\min}(J) \geq (1 - r) \min_j p_j$*. It is the Hessian of the convex potential* $s \mapsto \log\big(e^b + \sum_j e^{s_j}\big)$*. Moreover* $p = r\,\mathrm{softmax}(s)$ *with* $r = \sigma\big(\mathrm{LSE}(s) - b\big)$*.*

The inequality is Cauchy–Schwarz, $(\sum_j p_j x_j)^2 \leq r \sum_j p_j x_j^2$; the other statements are direct computations (Appendix A). So abstention and the convexity of log-sum-exp coexist: a head can learn to attend less, and its weights are still the gradient of a convex potential. For a numerical instance with $r = 0.953$, the smallest eigenvalue is 3.8×10^{-3}, above the bound $(1 - r) \min_j p_j = 3.1 \times 10^{-4}$; without the null it is zero to machine precision. The feedback gain of a loop through such a head scales with its row mass r as well; the monograph bounds that gain for heads whose rows sum to one and for heads that abstain (Bruce, 2026a, §§7.1–7.2).

There is also a norm picture. A softmax head over a sequence is a row-stochastic matrix P with a stationary law π, which is positive when the head is unmasked. Measure vectors by $\|v\|_\pi$, the square root of the π-weighted average of their squared components. Jensen's inequality with stationarity gives $\|Pv\|_\pi \leq \|v\|_\pi$ with equality on constants, so $\|P\|_\pi = 1$ exactly: the head can never shrink the constant direction. An abstaining head is a substochastic matrix Q, and when it is irreducible with at least one deficient row it is a strict contraction, $\|Q\|_w = \rho < 1$ in a weighted norm built from its Perron vectors (Theorem 3). Stochasticity forbids one more thing: a stochastic head cannot output zero for every collection of identical nonzero values. It can output zero by cancellation between values or through an output map, which is a different mechanism.

4 Gate times routing: every sink-logit or headwise-gated head is a diagonal gate on softmax

The mechanisms of Section 2 look different, but most of them are the same operation. Write $\mathrm{LSE}_i(s) = \log \sum_j e^{s_{ij}}$ for row i of a logit matrix s.

Proposition 2 (Gate times routing)**.** *A head with a null weight $z_i > 0$ in row i, $P_{ij} = e^{s_{ij}}/(z_i + \sum_k e^{s_{ik}})$, factors as*

$$P = \operatorname{diag}(g)\ \operatorname{softmax}_{\text{row}}(s), \qquad g_i = \sigma\big(\operatorname{LSE}_i(s) - b_i\big), \quad b_i = \log z_i.$$

Off-by-one softmax is $b_i = 0$; a learned sink is $b_i = b$, one trained number per head; value-null registers give a query-dependent b_i, the log of their summed exponential weight in row i.

A headwise output gate (Qiu et al., 2025), $o_i \leftarrow g_i o_i$ with $g_i = \sigma(w^\top x_i + c)$, has the same form with the gate driven by the query's input instead of its logits. So every sink-logit or headwise-gated head routes by an ordinary softmax and then scales each row by one number in $(0, 1)$: the gate decides how much the head says, the softmax decides where it looks. What differs between the mechanisms is only what drives the gate. Concurrent work (Wang, 2026) decomposes a gate on the values the same way, into a renormalized routing times a row attenuation (their Lemma 3.1). Sigmoid attention (Ramapuram et al., 2025), $P_{ij} = \sigma(s_{ij} + c)$, is outside the class: each key has its own gate, the row mass is not bounded by one, and nothing couples the keys. Its behaviour under the law is therefore different, and Proposition 6 treats it separately.

The Markov reading. An abstaining head is a killed Markov chain. Row i keeps mass $r_i = g_i$ and loses $1 - r_i$, so the head is a substochastic kernel Q. Conditioning on survival turns it back into attention. In the theorem, $D_x = \operatorname{diag}(x)$, $x \circ y$ is the entrywise product, and the reversal of a chain P with stationary law π is $P^*_{ij} = \pi_j P_{ji}/\pi_i$.

Theorem 3 (Killing, conditioning and reversal as attention)**.** *Let $Q \geq 0$ be nonzero, irreducible and substochastic with some row sum below one, with positive right and left Perron vectors h and ν and Perron root ρ. Then $0 < \rho < 1$, the conditioned kernel $P^h = \rho^{-1} D_h^{-1} Q D_h$ is stochastic with stationary law $\pi^h \propto \nu \circ h$, its reversal is $(P^h)^* = \rho^{-1} D_\nu^{-1} Q^\top D_\nu$, and $\|Q\|_w = \rho$ in the norm $\|x\|_w = \|D_h^{-1} x\|_{L^2(\pi^h)}$. If $Q_{ij} = e^{s_{ij} - L_i}$, then*

$$P^h = \operatorname{softmax}_{\text{row}}\big(s + \mathbf{1}(\log h)^\top\big), \qquad (P^h)^* = \operatorname{softmax}_{\text{row}}\big(s^\top + \mathbf{1}(\log \nu - L)^\top\big).$$

So the conditioned kernel P^h, the Doob transform of Q, is an ordinary softmax head with a key bias $\log h$, and its reversal is again an attention call. A path conditioned on survival to a finite time generally has a different law from a path of this kernel, as the proof of Proposition 12 shows. For a primitive Q, the endpoint conditioned on survival converges to $\nu/\nu^\top \mathbf{1}$, while the path occupation, the fraction of time a surviving path spends at each token, converges to π^h, in general a different law (Proposition 12). Primitivity rules out periodic oscillation of the endpoint.

The Perron pair is computable with log-sum-exp alone, since $\log(Qh)_i = \operatorname{LSE}_i\big(s + \mathbf{1}(\log h)^\top\big) - L_i$; in a numerical check on a 12-token off-by-one head, the corresponding iteration on the transposed logits gives ν and ρ to machine precision in 39 steps. The proof uses standard finite-state Perron and Markov-chain facts; we make no novelty claim for them, and the reversal machinery follows Fill (1991) and Bruce (2026a, §3.1).

Decoders are killed by construction. A causal mask supplies an absorbing state without any null. In a causal softmax head the first token attends only to itself, so e_1 is stationary and the spectrum is the diagonal. Tracking only the other tokens gives a killed chain: mass sent to the first token is lost from that chain. Every other row gives token 1 positive mass, so every transient token reaches absorption, the transient block T has $\|T^m\|_\infty < 1$ for some m, and repeated application drives all mass to token 1. The size of this structural sink is large in practice. A census of seven public causal language models at 64 tokens covers 2,700 captured operators, each averaged over 16 inputs: 2,544 heads stored in float16 and 156 head-averaged layers stored in float32. For each averaged operator, take the minimum mass sent to token 1 across its rows; this minimum is at least the mean of its per-input minima. Across operators its median is 0.416 (0.410 per head). The minimum falls below 10^{-6} in only 0.67% of operators, all of them heads; the smallest of these values are stored as exact zeros in float16. One caution: conditioned on survival, a causal chain concentrates on the transient token with the largest self-attention only if that token is reachable from where the chain starts. The chain $Q = \operatorname{diag}(0.9, 0.5)$ started in state 2 stays there whenever it survives.

5 The two-sided log-N law

5.1 Two matched queries share one null

Compare two matched queries. The *absent* query sees N background keys with scores $s_1, \dots, s_N$, total background weight $B_N = \sum_j e^{s_j}$, and a null of weight $z_N > 0$ (a single sink logit b has $z_N = e^b$; several value-null registers count through their summed weight). The *present* query sees the same background and null plus one matching key with score s_*. Abstention and retrieval are

$$a_N = \frac{z_N}{z_N + B_N}, \qquad w_N = \frac{e^{s_*}}{e^{s_*} + z_N + B_N}.$$

The matched convention is what makes the claim joint: the null cannot know whether the match is present. A gate that detects presence from outside the normalization is outside this class; Proposition 5 says what it can buy. Write $A_* = a_*/(1-a_*)$ and $W_* = w_*/(1-w_*)$ for the odds of two targets.

Theorem 4 (Exact simultaneous budget)**.** *For $0 < a_*, w_* < 1$, both $a_N \geq a_*$ and $w_N \geq w_*$ hold if and only if*

$$A_* B_N \;\leq\; z_N \;\leq\; e^{s_*}/W_* - B_N,$$

and such a null exists if and only if $e^{s_*} \geq w_* B_N / \big((1-w_*)(1-a_*)\big)$. If every content score lies in $[-\tau, \tau]$, no choice of null offset, fixed, learned, or a learned offset plus $\log N$, satisfies both beyond

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

and the bound is sharp: distractors at $-\tau$, the match at τ and $z_N = A_* N e^{-\tau}$ attain it.

Proof. Clearing the positive denominators turns each target into one side of the interval. The interval is nonempty exactly when $(1 + A_*)B_N \leq e^{s_*}/W_*$, which is the stated condition. With bounded scores, $B_N \geq N e^{-\tau}$ and $e^{s_*} \leq e^{\tau}$; the construction attains equality. □

The budget is one inequality with the null on both sides of it. Abstention needs the null to outweigh the background; retrieval needs the match to outweigh the null and the background together. Growing the background by a factor therefore squeezes the null's window from both ends, and with a bounded logit scale the window closes at N_*. The theorem decides any schedule, not only fixed scores: a schedule that multiplies the content scores by c_N and sets the null logit b_N keeps both targets at length N exactly when $A_* \sum_j e^{c_N s_j} \leq e^{b_N} \leq e^{c_N s_*}/W_* - \sum_j e^{c_N s_j}$. Its retrieval-only case, $a_* = 0$, is the bound on a softmax's largest weight of Nakanishi (2025, eqs. 5–6); see also Veličković et al. (2025, Lemma 2.1).

A null that sees presence. The budget binds because the matched queries share one null. A null that may differ between them by a bounded amount pays a known price.

Proposition 5 (Presence-sensitive null)**.** *Let the absent and present queries have background weights $B_0, B_1 > 0$, let the match have weight $M > 0$, and let their nulls have weights $z_0, z_1 > 0$ with $|\log z_0 - \log z_1| \leq \epsilon$. Both targets can be met if and only if*

$$M \;\geq\; W_*\big(B_1 + A_* e^{-\epsilon} B_0\big).$$

If every content score lies in $[-\tau, \tau]$, no such nulls meet both targets beyond

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

background keys, which rises from N_* at $\epsilon = 0$ to $N_*/(1-a_*)$ as $\epsilon \to \infty$, and the bound is sharp. If B and M are fixed with $A_* B > M/W_* - B$, a random null that meets the absent target with probability at least $1-\alpha$ and the present one with probability at least $1-\beta$ has absent and present laws that differ by at least $1-\alpha-\beta$ in total variation.

Sensitivity to presence can remove at most the abstention penalty, the factor $1 - a_*$ in N_*. It cannot remove the dilution of the match's weight by the background. It also needs evidence of presence: whatever the null is computed from must tell the two queries apart about as reliably as the targets demand.

5.2 What each rule does, and the mean-field law

A fixed sink, $z_N = e^b$, gives

$$a_N \le \sigma(b + \tau - \log N), \qquad w_N \le \frac{e^{2\tau}}{e^{2\tau} + N},$$

so both vanish; whenever the null scores below the match, as it must for the match to win, the abstention bound reaches any given level first, at $e^{b-\tau}$ times the length at which the retrieval bound does. A null calibrated to the length, $z_N = Ne^b$, holds abstention constant against a background of equal scores but gives $w_N = O(1/N)$. Sigmoid weights behave the same way.

Proposition 6 (Sigmoid weights)**.** *Let $p_j = \sigma(s_j + c_N)$ with one offset c_N shared by the matched queries, and define abstention, when $\sum_j p_j \le 1$, as $1 - \sum_j p_j$. If $|s_j|, |s_*| \le \tau$, the absent row mass is at most $M = 1 - a_*$, and the match weight is at least w_*, then necessarily $N \le M\big(1 + e^{2\tau}(1 - w_*)/w_*\big)$.*

A $-\log N$ bias is the choice $c_N = c - \log N$ that holds the row mass fixed; the proposition shows that it pays with single-key retrieval, whose weight then falls like $1/N$. None of these statements rules out perfect classification from a shrinking but nonzero signal in exact arithmetic: they bound attention weight.

The mean-field law. When the background exponentials e^{s_j} are i.i.d. with mean μ, abstention follows the mean-field prediction $a_{\mathrm{mf}} = z_N/(z_N + N\mu)$, and a fixed sink crosses one half at $N_{1/2} = e^b/\mu$.

Proposition 7 (Concentration and expectation error)**.** *If $e^{s_j} \in [l, u]$ with $l > 0$, then with probability at least $1 - \delta$,*

$$\frac{z_N}{z_N + N(\mu + t)} \le a_N \le \frac{z_N}{z_N + N\max(l, \mu - t)}, \qquad t = (u - l)\sqrt{\log(2/\delta)/(2N)},$$

and $0 \le \mathbb{E}a_N - a_{\mathrm{mf}} \le z_N N v/(z_N + Nl)^3$, where v is the variance of e^{s_j}.

For Gaussian logits $s \sim \mathcal{N}(m, v_s)$, $\mu = e^{m + v_s/2}$, and after tempering by λ the background moment is $e^{\lambda m + \lambda^2 v_s/2}$; the untempered μ is wrong after scaling. The moment predicts the realized background sum only while $\lambda\sqrt{v_s} < \sqrt{2\log N}$; past that scale the sum is carried by its few largest terms, and the moment overstates it by a factor of order $e^{(\lambda\sqrt{v_s} - \sqrt{2\log N})^2/2}$ or more.

A set of k near-matches contributes its exact sum separately. Heavy-tailed logits may have $\mathbb{E}e^s = \infty$ and then no finite mean-field law, while Theorem 4 still holds with the measured log-sum-exp.

In the pilot, the mean-field law, evaluated exactly with each head's learned logits, tracks the trained fixed-sink heads to within 0.002 per seed in all four configurations (Section 6).

5.3 Tempering every score, the null included

Suppose the head was trained at N_0 keys and is used at $N = N_0^\lambda$, that is $\lambda = \log N/\log N_0$, and multiply every score, the null included, by λ. The ratios in the theorem compare the largest background sum allowed at the training length with the null and with the match, and the null with the match.

Theorem 8 (Sufficient separated margins)**.** *Let the background scores be at most s_0, with $s_* > b > s_0$ fixed before scaling, and put $v_a = N_0 e^{s_0 - b}$, $v_r = N_0 e^{s_0 - s_*}$ and $u_r = e^{b - s_*}$. After tempering,*

$$a_N \ge \big(1 + v_a^\lambda\big)^{-1}, \qquad w_N \ge \big(1 + u_r^\lambda + v_r^\lambda\big)^{-1}.$$

If $v_a, v_r, u_r \le 1$, both lower bounds are non-decreasing in N; if all three are below one, both tend to one.

Since $s_* > b$, $u_r < 1$ and $v_r = u_r v_a < v_a$, so the single condition $v_a \le 1$ suffices. The margins are real conditions. If the null only equals the level of a uniform background, $b = s_0$, tempering changes no relative score of the absent query and $a_N = 1/(N+1)$ at every λ; more generally $b - s_0 < \log N_0$ makes the abstention bound decay even when the null outranks every single distractor. So a growing logit budget is necessary in this class (Theorem 4), and a particular schedule is not sufficient without the margins. With one common multiplier, abstention depends exactly on a power mean of the exponentiated background scores. This average lies between their ordinary mean and their maximum:

Theorem 9 (Power-mean law). *For an integer $N \geq N_0$, $\lambda = \log N / \log N_0$, and arbitrary finite scores, let $M_{q,N} = \left(N^{-1} \sum_j e^{q s_j}\right)^{1/q}$. Then exactly*

$$a_N = \left[1 + \left(N_0 e^{-b} M_{\lambda,N}\right)^{\lambda}\right]^{-1}, \qquad \frac{1}{1 + \left(N_0 e^{s_{\max} - b}\right)^{\lambda}} \leq a_N \leq \frac{1}{1 + \left(N_0 e^{-b} M_{1,N}\right)^{\lambda}}.$$

The upper bound, which uses the ordinary mean $M_{1,N}$, can never certify a target; the lower bound is the separated-margin bound of Theorem 8. The law needs the same multiplier on every score. YaRN multiplies every content score by a temperature (Peng et al., 2024); a sink logit left out of that factor breaks the premise, and in an explicit example tempering then loses abstention (0.62 to 0.37 from one to eight times the training length) where the common multiplier keeps it (0.62 to 0.66); the example has $N_0 = 1{,}024$ background logits at $s_0 = 2$, the null set so that $v_a = e^{-0.5}$, and YaRN's factor $(0.1 \ln 8 + 1)^2$ at eight times. The trained harness gave its tempered sinks YaRN's factor as well, so there the premise held (Section 8). For a random background the law has a sharp phase boundary.

Theorem 10 (Almost-sure converse). *Let the background scores be i.i.d. with essential supremum s_0 and $p_0 = \mathbb{P}(S = s_0)$, and let $h = \log N_0 + s_0$. Under tempering, almost surely*

$$a_N \longrightarrow \begin{cases} 0, & b < h, \\ (1 + p_0)^{-1}, & b = h, \\ 1, & b > h. \end{cases}$$

A present match with score s_* keeps weight tending to zero if $s_* < \max(b, h)$, and to $\left[1 + \mathbf{1}_{\{s_* = b\}} + p_0 \mathbf{1}_{\{s_* = h\}}\right]^{-1}$ otherwise. If the score law has unbounded upper support, $a_N \to 0$ for every finite b.

So the abstention margin $v_a \leq 1$ is necessary as well as sufficient, up to the boundary. The boundary depends on the atom p_0 at the top of the score law, meaning the probability of a score exactly at that top. Here s_0 must be the essential supremum, the smallest upper bound that holds with probability one, not a loose bound. The same conclusions hold for stationary ergodic backgrounds; a nonergodic one need not use its unconditional supremum. At the boundary the rate of convergence is set by the upper tail, the probability of scores close to the top:

Theorem 11 (Boundary rate). *If in addition $\mathbb{P}(s_0 - S \leq t) \sim C t^a$ as $t \downarrow 0$, with $C > 0$ and $a > 0$, then at $v_a = 1$, almost surely $1 - a_N \sim C\,\Gamma(a+1)\lambda^{-a}$, and a null lowered by $\left(a \log \lambda - \log(C\Gamma(a+1)) + t\right)/\lambda$ below the boundary holds abstention at $(1 + e^t)^{-1}$.*

These results concern attention weights under a fixed score law. They do not identify the law a network learns as the context changes, and they do not turn an accuracy onset into a weight threshold (Appendix A has the proofs and a finite-sample version of Theorem 10).

Weight is not accuracy. The theorems bound weight. The retrieval models of Section 10 measure accuracy under a learned decoder on noiseless values instead. Their native onsets, the first lengths at which that accuracy falls below one half, are exactly 8,192, 32,768 and 131,072 keys at training lengths 32, 64 and 128, that is $8N_0^2$, in every seed, while the trained logit gap g_0 rises from about 4.2 to 5.5. A line through these three groups gives $\log_2(\text{onset}) = 3.085\, g_0 + 0.003$, an onset growing as $e^{2.14\, g_0}$, faster than the weight bound's e^{g_0}, though the design cannot separate g_0 from the training length. A learned decoder scored by nearest cosine can read a match whose weight has already fallen, so the accuracy onset is a separate estimand (Section 10).

5.4 Compensation through position

A trained network can escape a fixed sink's decay by lowering its background logits as the context grows. The sink's log-odds against the background, $y_i = b - \mathrm{LSE}_j(s_{ij})$, fall with slope $-\gamma$ in $\log i$, where γ is how fast the background grows; a stationary background with a finite exponential moment has $\gamma = 1$. In small trained models without positional encoding and with a learned null, the median within-model slope, pooled

over the heads of 5 seeds at each training length, is −1.04 to −1.01 in the models trained at more than 64 tokens and −1.31 at 64 tokens (Section 7).

With rotary positions the kernel holds the background down, and training widens it only when the null itself grows with length. The 125M results summarized under this heading average each model's median head slope over 3 seeds per rule. On FineWeb-Edu text inside the training length, the background grows like $n^{0.12}$ for a fixed null, whose log-odds then fall at only −0.12, and like $n^{1.07}$ for a null that adds $\ln n$, so abstention stays nearly level under either rule (Section 8).

Small byte models show the same contrast with the positions manipulated: γ is 0.12 to 0.19 with rotary positions and no null, an output gate or a fixed null, 0.97 with a null that adds $\ln n$, and 1.08 to 1.13 without positions whatever the null (Section 7).

The compensation works through position and holds while positions stay in the trained range: beyond it the fixed sinks' slopes are −0.29 and −0.28 with positions interpolated (YaRN), and about −6.2 natively.

gpt-oss has learned sink logits in dense layers and in 128-token banded layers, whose background stops growing at 128 keys, a built-in control. A prespecified test of its slopes was undecided by its rule: its books control failed, and the measured slopes match neither prespecified branch. Descriptively, they lean toward compensation through position (Section 9).

The form of the compensation. Each head's content logits split exactly into a level r_i for each query, a popularity c_j for each key and a cycle part C_{ij}, the remainder orthogonal to every such sum: $s_{ij} = r_i + c_j + C_{ij}$, fitted by least squares over the sampled query rows and every key each one sees. The level and popularity together form the gauge part. Then $\mathrm{LSE}_j(s_{ij}) = r_i + \mathrm{LSE}_j(c_j) + X_i$ exactly, where X_i is the log of $e^{C_{ij}}$ averaged under the popularity weights, so the background's growth is a level term plus a popularity term plus a cycle term. A softmax over the keys cannot see the level, which cancels from its weights; the null's log-odds y_i see all three terms.

In the 125M language models with a fixed null, inside the training length, the level's slope is −1.55 per e-fold of visible keys, the popularity term's 1.86 and the cycle term's −0.003 (95% interval [−0.012, 0.011]; medians over 432 heads, on eight PG-19 books). The background itself grows like $n^{0.25}$ ([0.20, 0.34]) on these PG-19 books, against $n^{0.12}$ on FineWeb-Edu text; medians of the components do not add. The pooled estimates here use 3 seeds at 125M and four scored seeds in the byte models; their intervals are 95% cluster bootstrap intervals, resampling seeds and then heads within each seed.

So the compensation lives in the gauge part, which includes the linear part of a recency kernel: a bias linear in $i - j$, as in ALiBi (Press et al., 2022), is exactly a level falling with i plus a popularity rising with j, with no cycle part. The softmax never sees the falling level, but against the null it offsets most of the popularity term's growth. The small byte models with rotary positions and a fixed null show the same pair (−1.49 and 2.49) beside a real cycle term (−0.70, [−0.85, −0.57]); without positions the level is flat (−0.02) and the popularity term grows at 1.11 to 1.12, near the rate of 1 for keys of equal popularity.

A null that adds $\ln n$ flattens the level's fall, by 1.12 per e-fold at 125M ([0.95, 1.27]) and 1.22 in the byte models: each model sets its gauge to the null it has.

Beyond the training length the level reverses natively. On one split fitted over every length, its slope beyond the training length exceeds its slope inside it by 7.03 per e-fold at 125M ([6.53, 7.48]) and 1.89 in the byte models.

Interpolated positions hold the background instead. At 125M with a fixed null, beyond the training length, the background grows like $n^{0.28}$ to $n^{0.35}$ across seeds under interpolation, against $n^{6.5}$ to $n^{7.3}$ natively, and interpolation holds it by steepening the gauge: the level falls faster and the popularity term grows faster than within the training length. A prespecified prediction that interpolation would keep the level's slope within 0.6 of its in-range value therefore failed at 125M (a shift of 1.55, [1.41, 1.66]), while it held in the byte models (a shift of 0.31).

5.5 A law of sums, and two limits

Every rule in this section decides by comparing one logit with a sum over the N keys: softmax, sinks, registers, gated heads and sigmoid weights. A head that decides by reading a maximum, abstaining unless its best score clears a threshold, needs no schedule against backgrounds bounded below its threshold, and otherwise only a threshold that tracks the background's extreme value; it fails only when the background maximum reaches the threshold (Bruce, 2026e). A softmax whose scale grows past the point where the background sum is carried by its largest terms reads the maximum in the same way. The harness included a head that reads a maximum: with YaRN it was the only rule whose null and match weights both changed by less than half a nat of log-odds out to eight times the training length, on the seed means though not in every seed, and without YaRN it failed as the matching key fell far down the ranking and the background maximum grew with length (Section 8).

Two limits. The budget of Theorem 4 is a limit of the read, not of the keys. A soft rule whose log-odds between two keys are capped at 2τ cannot certify more than about $e^{2\tau}$ alternatives, whatever the keys are; tempering raises the cap, and a head that reads a maximum has none. A second limit belongs to the keys. Random keys in $\mathbb{R}^d$ form a spherical code, and a match at cosine c with the query falls behind the largest of N background keys, with probability one half, at

$$N_{\text{ev}} \approx \ln 2 \; c\sqrt{2\pi(d-1)}\; e^{(d-1)C}, \qquad C = \tfrac{1}{2}\ln\Big(1 + \frac{c^2}{1-c^2}\Big),$$

where C is the capacity of a Gaussian channel at the signal-to-noise ratio $c^2/(1-c^2)$ (Shannon, 1959). A rule that reads these keys by their scores, a sum whose weights rise with the score or a maximum, puts most of its weight on the match, or selects it, only when the match has the highest score (Bruce, 2026e), so none retrieves with probability above one half beyond it. When the match is equally likely to be any key and its own score is random, with independent background scores and a likelihood ratio that rises with the score, no rule that sees only the scores finds the match more often than picking the highest score does (Bruce, 2026e). So tempering lifts a sum rule toward this ceiling and never past it, as random-energy analyses of softmax retrieval also find (Lucibello & Mézard, 2024).

When the match's score is noisy and its log-likelihood ratio is linear in the score, the posterior over 'no match' and each key is a softmax at the true noise level whose null logit is $\log N$ plus a constant, and thresholding it is the optimal decision with an abstention option (Forney, 1968). With noise-free matches, as in the pilot, there is no finite true temperature: the best read that never ranks a lower score above a higher one is the maximum, which tempering approaches from below. Either way, what fails at long context is a fixed temperature and a fixed null, not the summing.

In the pilot the matches are exact ($c = 1$), the code is noiseless, and only the first limit binds.

A direct test with explicit random keys confirms both limits (Bruce, 2026e). In each of 20 settings the argmax's onset follows the exact law, and it grows with d at the rate C, not at its small-c form $c^2/2$. Raising τ from 2.5 to 80 moves softmax's onset from 3.4–10 nats below that wall to within 0.02–0.81 nats of it, and never above. The head dimension is therefore part of the routing budget: at $d = 128$, the strongest of 10^6 random keys sits at cosine 0.41, so a weaker match is lost even to a rule that reads the maximum.

Trained heads show the same split (Bruce, 2026e): with the match at cosine 0.6 or 0.8 among four hard distractors and d between 11 and 38, a head that reads the maximum and a tempered sum retrieve up to the code limit at every learned scale, tracking its slope in d to within 4% of C, while a sum at a fixed scale falls further below it as d grows.

6 A controlled pilot: seven rules, two budgets, two training lengths

The laws of Section 5 describe one head with given scores. The pilot trains such heads and asks whether they still behave as the laws say once their scale, sink and register are learned. Each head performs lookup-or-abstain. A query is compared with N keys by cosine similarity, with queries and keys isotropic in $\mathbb{R}^{128}$. Half the queries have an exact match, and the head must return its value, one of 16 one-hot classes; the other half

	Softmax	Value-null register	Fixed sink	Sink $+\log N$	Sigmoid	Sigmoid $-\log N$	$\log N$-tempered
$\tau_{\max} = 10$, $N_{\text{train}} = 64$							
learned τ	7.6	9.8	9.8	9.7	6.9	9.8	9.8
abstention fails	$\leq$16	128/256	512	—	64	—	—
retrieval fails	—	—	—	2,048	8,192	2,048	—
$\tau_{\max} = 10$, $N_{\text{train}} = 512$							
learned τ	9.6	9.8	9.8	9.8	1.1	9.9	9.8
abstention fails	$\leq$16	256	1,024	—	64	—	—
retrieval fails	—	—	—	32,768	$\leq$16	16,384	—
$\tau_{\max} = 20$, $N_{\text{train}} = 64$							
learned τ	7.6	16.7/16.6	15.2	15.6	6.9	12.9	15.2
abstention fails	$\leq$16	2,048	8,192	—	64	—	—
retrieval fails	—	—	—	32,768	8,192	16,384	—
$\tau_{\max} = 20$, $N_{\text{train}} = 512$							
learned τ	10.6	19.6	17.8	17.8	2.0	15.0/15.1	17.8
abstention fails	$\leq$16	8,192/16,384	65,536/32,768	—	64	—	—
retrieval fails	—	—	—	—	$\leq$16	—	—

Table 1: Pilot accuracy onsets: the first length N at which abstention accuracy (absent queries) or retrieval accuracy (present queries) falls below 0.9, with the learned logit scale τ. Entries show seeds 0 and 1 as "seed 0/seed 1" where they differ; — means no onset up to 65,536 keys, and $\leq$16 means below 0.9 already at the shortest length. Weight onsets, which can differ, are in Table 8. The tempered head is the fixed sink's trained head, tempered at evaluation.

have none, and the head must return zero. The output is $o = \sum_j w_j v_j$, trained by squared error, and the logits are $s_j = \tau \cos_j + \beta_0$ with a learned scale $\tau = \tau_{\max}\sigma(\theta)$ bounded by a budget $\tau_{\max} \in \{10, 20\}$. Seven rules share this geometry:

- **softmax**, which cannot abstain;
- a **value-null register**: an extra key with zero value and logit $\tau \cos(q, k_{\text{reg}}) + \beta_0 + b$, whose direction is random;
- a **fixed sink**: a learned null logit b, the form gpt-oss uses (OpenAI, 2025);
- **sink** $+\log N$: the sink logit raised to $b + \log N$, calibrated to the length;
- **sigmoid**: $w_j = \sigma(s_j + b)$, with no normalization;
- **sigmoid** $-\log N$: $w_j = \sigma(s_j + b - \log N)$, the calibrated recipe of Ramapuram et al. (2025);
- $\log N$**-tempered**: the fixed sink with every score, the sink included, multiplied by $\lambda_N = \log N / \log N_{\text{train}}$.

Each rule is trained at $N_{\text{train}} \in \{64, 512\}$ keys under each budget, with two seeds, and evaluated in float64 at every power of two from 16 to 65,536 keys (Appendix B; all four configurations are in Figure 12 and Table 9). Since $\lambda_N = 1$ at the training length, the tempered head trains exactly as the fixed sink does and has the same parameters. Tempering acts only at evaluation, so in this pilot it costs no training.

A head abstains on a query when its row mass $\sum_j w_j$ is below 1/2. It retrieves correctly when the mass is at least 1/2 and the match's class is the largest entry of o. We report two estimands and keep them apart. The weights are the abstention $a = 1 - \sum_j w_j$ on absent queries and the match weight w on present ones; a weight's onset is the first length at which its mean falls below 1/2. The accuracies are the fractions of absent queries that abstain and of present queries that retrieve; an accuracy's onset is the first length at which it falls below 0.9 (Table 1).

6.1 Each fixed rule fails at least one side

At $\tau_{\max} = 10$ and $N_{\text{train}} = 64$ (Figure 2), the fixed sink keeps retrieval and loses abstention. Its absent-query abstention falls from 0.874 at the training length to 0.097 at 4,096 keys and 0.007 at 65,536 (seed means),

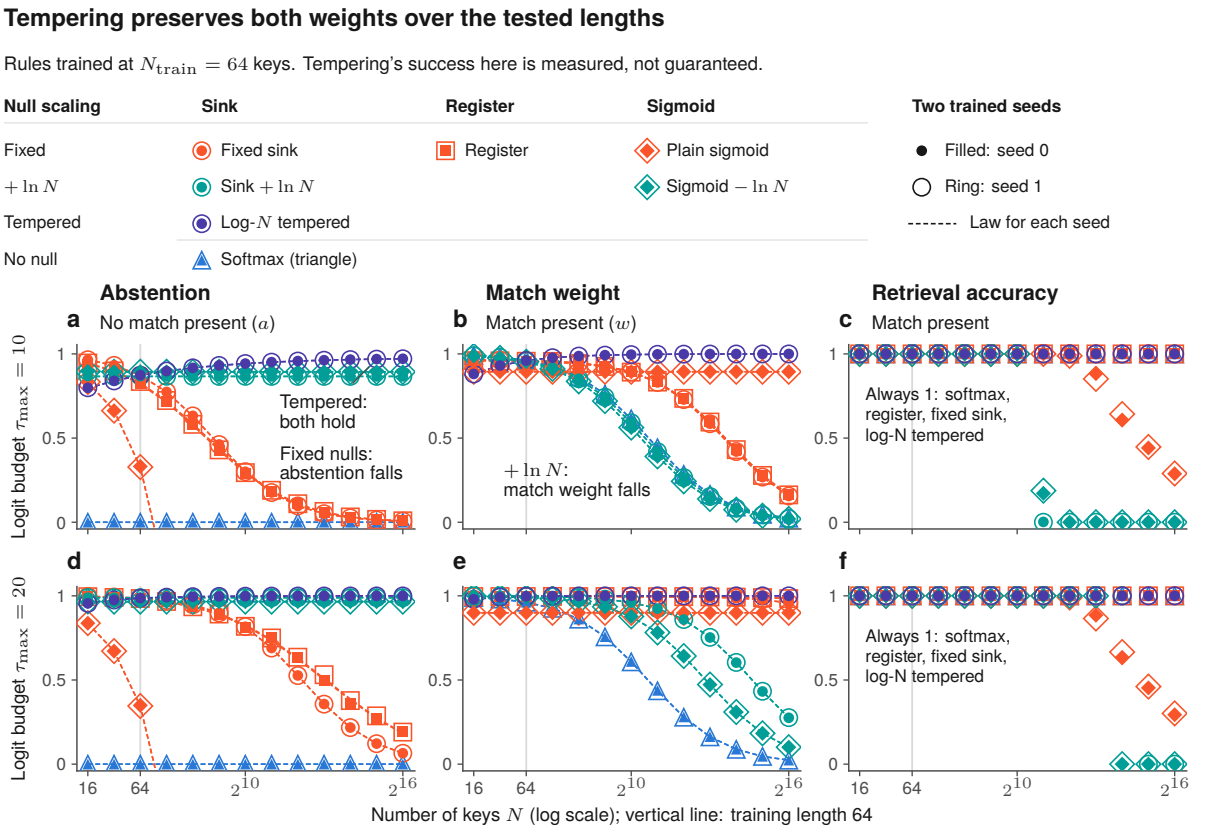


Figure 2: Tempering keeps both weights over the tested lengths; every other rule loses one of them. Seven attention rules, trained at 64 keys, at two logit budgets (rows). Colour says how a rule's null scales with the number of keys N: fixed (orange), raised by $\ln N$ (teal) or tempered together with every score (violet); softmax (blue) has no null. Shape says the mechanism: a sink (circle), a register (square) or a sigmoid (diamond); softmax is a triangle. (a, d) Abstention on queries with no match, $a = 1 - \sum_j w_j$: the fixed nulls lose it as N grows. (b, e) The match weight w on queries with a match: the nulls raised by $\ln N$ lose it. (c, f) Retrieval accuracy, which also requires the row's total weight to be at least one half; softmax, the register, the fixed sink and tempering stay at one throughout. A filled mark and a ring are seeds 0 and 1; dashed lines are the law's prediction for each seed's trained head. Plain sigmoid's abstention turns negative from $N = 128$, below the plotted range. Tempering's success here is measured, not guaranteed: its trained margins over the background do not meet the conditions of Theorem 8. The pilot uses isotropic keys in $\mathbb{R}^{128}$, exact matches and 1,024 evaluation queries per length and seed (Appendix B).

and its abstention accuracy fails from 512 keys in both seeds. The value-null register fails the same way, from 128 keys in seed 0 and 256 in seed 1. Both retrieve correctly at every length.

At this setting, the calibrated rules do the opposite. Sink $+\log N$ and sigmoid $-\log N$ hold their abstention near its training value (0.866 and 0.893 at 65,536 keys) but lose retrieval from 2,048 keys in both seeds, where the match weight has fallen to 0.42 and 0.39.

Softmax never abstains. Plain sigmoid cannot hold its row mass: the mass exceeds one from 128 keys in every configuration and seed, and at $\tau_{\max} = 10$, $N_{\text{train}} = 64$ it reaches 689 at 65,536 keys, where its retrieval accuracy has fallen to 0.29 (from 0.99 at 4,096). At $N_{\text{train}} = 512$ it does not learn to retrieve at its training length: its learned τ is 1.1 and 2.0 under the two budgets, and its retrieval accuracy there is 0.12 and 0.31.

The tempered rule keeps both sides. At $\tau_{\max} = 10$, $N_{\text{train}} = 64$ its abstention does not fall (0.958 at 4,096 keys and 0.971 at 65,536), and none of its weights or accuracies reaches an onset at any length, in any configuration or seed.

A larger budget delays the failures but keeps the trade. Doubling the budget to $\tau_{\max} = 20$ at $N_{\text{train}} = 64$ moves the fixed sink's abstention failure to 8,192 keys, and the calibrated rules' retrieval failures to 16,384 keys (sigmoid $-\log N$) and 32,768 (sink $+\log N$). Training at 512 keys as well moves them further. The calibrated rules then retrieve to 65,536 keys. The fixed sink's abstention accuracy first fails at 65,536 keys in seed 0 and in the seed mean, while seed 1 dips to 0.898 at 32,768 keys; its abstention weight falls below 1/2 at 65,536 keys in both seeds. Only the rules that use the budget move. Softmax and sigmoid learn $\tau \approx 7.6$ and $\tau \approx 6.9$ at $N_{\text{train}} = 64$ under either budget, and their onsets there do not change.

Accuracy and weight are different estimands. The two onsets need not agree, and the pilot shows both directions. Softmax retrieves correctly at every length, although at $\tau_{\max} = 10$ and $N_{\text{train}} = 64$ its match weight falls below 1/2 from 2,048 keys, to 0.024 at 65,536: a correct read needs only that the match's class lead the others, and the background spreads its weight over 16 classes. Conversely, at $\tau_{\max} = 20$ and $N_{\text{train}} = 512$ the register's mean abstention stays above 1/2 to 65,536 keys (0.541 there), while its abstention accuracy fails from 8,192 keys in seed 0 and 16,384 in seed 1. The mean hides a growing share of rows that no longer abstain.

6.2 The mean-field law and tempering in the trained heads

The law tracks the trained heads. The mean-field law of Section 5, with its expectation computed exactly for isotropic keys and each head's learned parameters, and with nothing fitted, tracks the trained fixed-sink heads to within 0.002 per seed in abstention (largest error 0.0017) and 0.00022 in match weight, in all four configurations. The six untempered rules have 96 weight sides in all, two for each combination of rule, configuration and seed: abstention and match weight. On the 64 with a measured onset, the measured onset and the law's agree to within one grid step. On the other 32, the law places no onset on the grid either (Figure 3, Table 8).

Tempering works here without meeting its certificate. The tempered rule's trained margins do not satisfy Theorem 8. Measured against the background supremum, a key at cosine one, the absent margin is $v_a = 1{,}777$ and 1,779 in the two seeds at $\tau_{\max} = 10$, $N_{\text{train}} = 64$, and between 7,304 and 241,450 in the other configurations; the present margin v_r equals N_{train}, because a background key at the supremum would score as high as the match. Per row the margins look better: at the training length, the median row's largest background score gives $v_a = 0.73$ at $\tau_{\max} = 10$, $N_{\text{train}} = 64$. But the theorem needs the bound fixed before scaling, and the typical row's background maximum rises with N, by 1.6 logits between 64 and 65,536 keys in that configuration, so the per-row reading does not transfer. What describes the tempered heads is the power-mean law of Theorem 9, with the background's population power mean in place of its sample one. We check this stationary law numerically and do not prove it: evaluated with each head's learned parameters, it reproduces the measured tempered abstention to within 0.0023 over the grid. It predicts that tempered abstention falls below 1/2 near $N = 2.3 \times 10^{12}$ keys at $\tau_{\max} = 10$, $N_{\text{train}} = 64$, near 8.4–8.6×10^{6} at $\tau_{\max} = 10$, $N_{\text{train}} = 512$, near 9.3–9.4×10^{12} at $\tau_{\max} = 20$, $N_{\text{train}} = 64$, and near 1.5×10^{14} at $\tau_{\max} = 20$, $N_{\text{train}} = 512$ (Table 2). These lengths extrapolate a law checked only to 65,536 keys; they are not measurements.

6.3 Other key geometries and training lengths

Scope. The pilot fixes the feature geometry: keys are isotropic and fixed, matches are exact, and only the scale τ, the offset β_0 and the null logit b are learned. A trained network can also move its background logits with N, which this pilot rules out by construction; gpt-oss's background falls with position (Section 9). A registered extension relaxes the geometry, the values and the training length, with the pilot's training and evaluation and five seeds per setting.

The same heads are trained at both budgets and $N_{\text{train}} = 64$ in the pilot's geometry and seven others (Appendix B, Table 10): keys drawn around 16 clusters, background keys with a heavy-tailed component

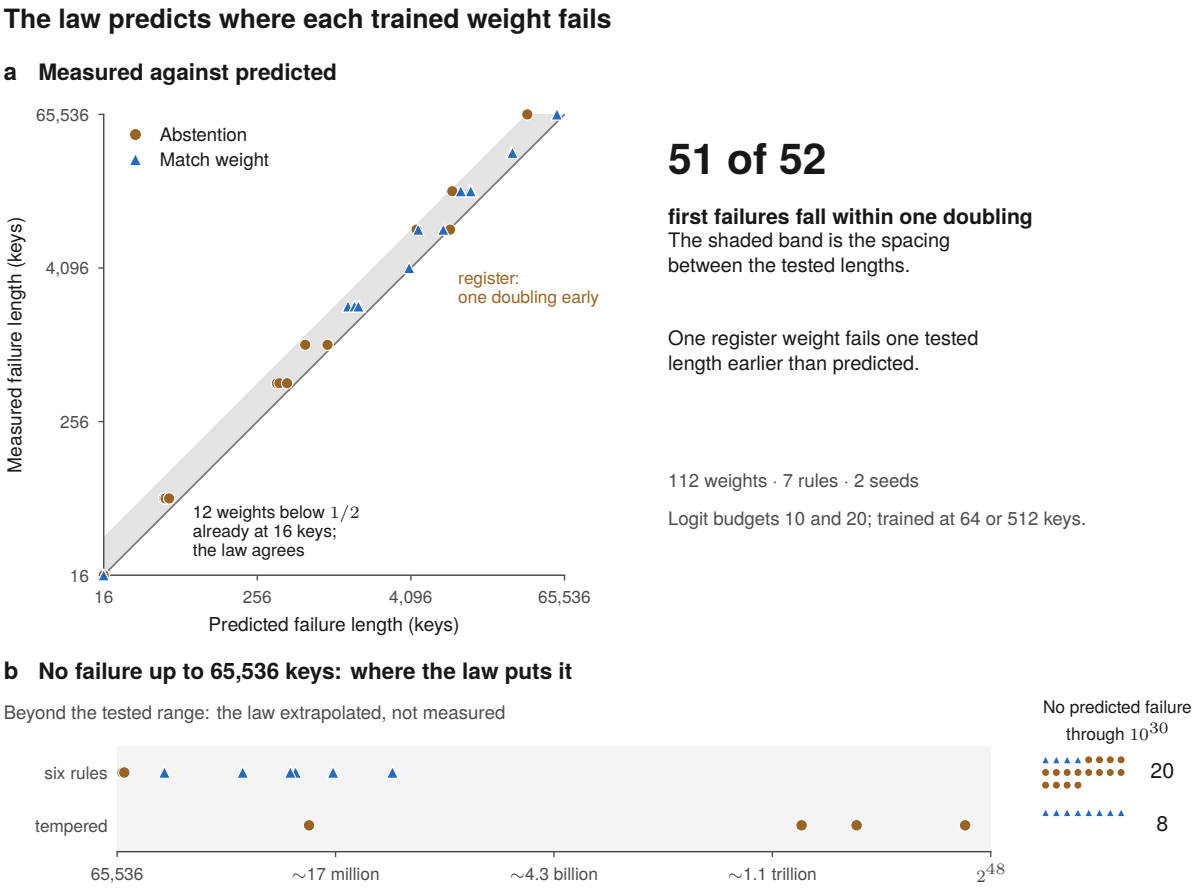


Figure 3: The mean-field law predicts where each trained head's weights fail. (a) For the six untempered rules, measured failure length against the prediction from each head's learned parameters, with nothing fitted: abstention on absent queries (brown circles) and match weight on present queries (blue triangles). Of the 52 weights that first fall below one half after the shortest tested length, 51 fall between the prediction and twice it (shaded), the resolution of the evaluated lengths. The remaining register weight fails one doubling before the first evaluated length at or above its prediction. Another 12 weights are already below one half at 16 keys, as the law predicts. (b) The 48 weights with no measured failure through 65,536 keys: the law predicts a later failure for 12 of the six untempered rules' weights (about $2^{16.2}$–$2^{26.1}$ keys) and all 8 tempered abstention weights (about $2^{23.0}$–$2^{47.1}$ keys). For the remaining 28, including all 8 tempered match weights, the law has no failure through 10^{30} keys; this is not an infinite-range guarantee. These are extrapolations of laws checked only to 65,536 keys, not measurements. The tempered prediction uses the numerically checked stationary power-mean law and does not certify the trained margins over the background. All 112 weights are represented: seven rules, logit budgets 10 and 20, training lengths 64 and 512, and 2 seeds, with 1,024 evaluation queries per length and seed. The shading is not a confidence interval; the quantities are mean attention weights, not task accuracies.

				Supremum margins			Per row			
$(\tau_{\max}, N_{\text{train}})$	Seed	τ	$b - \beta_0$	v_a	v_r	u_r	median v_a	drift	Law error	Failure N
(10, 64)	0	9.78	6.46	1,777	64	0.036	0.73	1.59	0.0018	2.3×10^{12}
	1	9.78	6.46	1,779	64	0.036	0.73	1.59	0.0023	2.3×10^{12}
(10, 512)	0	9.85	7.19	7,314	512	0.070	5.10	1.01	0.0022	8.4×10^{6}
	1	9.85	7.19	7,304	512	0.070	5.10	1.01	0.0022	8.6×10^{6}
(20, 64)	0	15.18	9.32	22,449	64	0.003	0.12	2.47	0.0007	9.4×10^{12}
	1	15.17	9.31	22,439	64	0.003	0.12	2.47	0.0010	9.3×10^{12}
(20, 512)	0	17.80	11.64	241,020	512	0.002	0.48	1.83	0.0012	1.5×10^{14}
	1	17.80	11.65	241,450	512	0.002	0.48	1.83	0.0015	1.5×10^{14}

Table 2: The tempered head at its trained parameters. The supremum margins v_a, v_r and u_r of Theorem 8 use the background supremum, a key at cosine one; the theorem certifies the head only if all three are at most one. The per-row columns give the median row's v_a at the training length, from its largest background score, and the drift: how far that maximum rises, in logits, between N_{train} and 65,536 keys. The law error is the largest gap between the measured abstention and the stationary power-mean law over the grid. Failure N is where that law puts abstention below 1/2, an extrapolation beyond the grid.

along the query, a noisy match at cosine 0.95, 0.80 or 0.60 among four hard distractors, and continuous values read by one head or by four 32-dimensional heads. The rules that sum over the keys keep the pilot's pattern: none without a length input keeps both sides where the law puts a failure inside the grid. The registered outcome is nonetheless partial, because the pattern needs two qualifications, and with near matches at cosine 0.80 and 0.95 at $\tau_{\max} = 20$ the register's abstention is neither kept nor lost by the registered thresholds in three cells (Appendix B).

With heavy-tailed keys every rule that sums over the keys loses a side, tempering included. The calibrated rules lose retrieval, as predicted, and keep their abstention accuracy, but their mean abstention also falls with N, by 0.034 to 0.058, beyond the registered tolerance of 0.02. Tempering loses abstention accuracy from 256 to 1,024 keys. Its own law, given the heavy-tailed background, also puts the mean abstention below 1/2 inside the grid, at 370 to 600 keys, though earlier than the measured mean falls below it and outside the regime in which the law is quantitative.

With clustered keys at $\tau_{\max} = 10$, the tempered heads' mean abstention stays above 0.64 to 65,536 keys, and the law puts its fall below 1/2 beyond the grid; yet their abstention accuracy falls below 0.9 inside the grid, to between 0.71 and 0.77 at 65,536 keys, in all five seeds.

Training on many lengths. With isotropic keys, training on lengths drawn from 64 to 1,024 delays the fixed sink's failure without removing it: its abstention accuracy fails from 1,024 keys instead of 512 at $\tau_{\max} = 10$, and from 32,768 instead of 8,192 at $\tau_{\max} = 20$; with clustered or heavy-tailed keys at $\tau_{\max} = 10$ it does not delay the failure (Table 11).

A sink whose null logit adds $\gamma \log(N/64)$, with γ learned, keeps both sides only at the larger budget: in all five seeds with isotropic keys, where it learns $\hat{\gamma} = 0.95$ to 1.06, a $\log N$ schedule whose match weight the law puts below 1/2 at or just past the end of the grid, and in two with clustered keys ($\hat{\gamma} = 0.63$ to 0.66). At the smaller budget it learns $\hat{\gamma} \leq 0.31$ and loses abstention, and with heavy-tailed keys it loses both sides.

Trained on the same lengths with heavy-tailed keys, the calibrated sink's mean abstention falls by 0.09 to 0.22 over the grid, and at the smaller budget it loses abstention in three of five seeds.

Tempering applied in training loses abstention with heavy-tailed keys, from 512 to 2,048 keys, and with clustered keys at the smaller budget in four of five seeds.

The law in other geometries. The mean-field law, evaluated with each geometry's own background moment, agrees with the measured weight onset in 1,113 of the 1,129 scored comparisons. For onsets inside

the grid, agreement allows one grid step in N from the first grid point at or above the predicted onset; censored onsets have separate comparison rules (Appendix B). All 16 misses are the register with near matches (Figure 13); another 1,151 comparisons are not scored, because the law is outside its concentration regime there or neither onset falls on the grid. Every geometry passes at least 91% of its comparisons over both budgets together, but at $\tau_{\max} = 20$ near matches at cosine 0.80 and 0.95 pass only 20 of 25 and 25 of 29, below the registered 90%, so there the mean-field law is not quantitative.

Tempering is certified in no geometry. Of the 120 labelled cases, 114 have their abstention tracked by the stationary law of a concentrated background to within 0.05 at every length. The remaining six are empirical successes: they keep both sides without that explanation. With heavy-tailed keys tempering has no label, because it fails. The law also tracks cases that fail. At the smaller budget 29 of those 114 cases lose abstention: in the 20 with four 32-dimensional heads the law puts the failure inside the grid, and in the rest, with clustered keys, it tracks a mean weight that stays above 0.64, while the failure is in accuracy, which the law does not predict. In 15 more, with near matches, the head did not abstain even at the training length.

Two heads that decide by reading the largest score are scored separately. The key check, which reads it for both decisions, keeps both sides in every geometry except heavy-tailed keys, where it loses both, its abstention failing at the wall its extreme-value law predicts; the max gate, which reads it to abstain but routes over every key, also loses both with heavy-tailed keys and keeps both in the other geometries, except that it loses retrieval with the weakest near match at the smaller budget. The need for a budget that grows like $\log N$ is therefore a property of rules that sum over the keys (Section 5; Bruce, 2026e).

7 What trained models already show

We trained small models that each change one thing about how a head attends, and measured the sinks, the massive activations and the quality that result (Table 3). We read them against Section 3: a head given a way to attend less should need its sinks less. The models are small and the seeds few. Where a comparison has five seeds we give the paired change with its 95% interval; with two or three seeds we give the means and the number of seeds that moved in the stated direction, and every seed's value is in Table 12. We give no p-values.

7.1 The readouts, and two ways to attend less

The readouts. Massive activation is the largest hidden-state magnitude divided by the median one, per layer, then the maximum over layers (for sigmoid attention and the four registers) or the mean (for the output gate and the value-null register). In the bidirectional ViTs a sink is a token with high stationary mass and a small value norm; the detector's third test, a high loading on the constant mode, ranks tokens exactly as stationary mass does there. In the causal models the stationary test is vacuous: the first token absorbs the chain, every other token's stationary mass sits at the numerical floor, and every token passes. The loading test is numerically degenerate there as well, yet its top-decile cut still selects. So a causal model's content-sink mass rests on the value-norm test and on that degenerate loading: read it as attention to a detector-selected subset of tokens with small value norms, not as mass on stationary-law sinks. First-token mass, where recorded, is read directly and uses no stationary law.

Two ways to attend less remove much of the sink mass. Replacing softmax with sigmoid attention in a CIFAR-100 ViT, with unnormalized scores and the output scaled by $1/N$, lets rows abstain (its sink readouts use the row-renormalized attention): the mean row mass falls to 0.34 at depth 12 and 0.31 at depth 24. Content-sink mass falls from 0.026 to 0.002 and from 0.033 to 0.001 (lower in 3/3 seeds at depth 12 and 3/3 at depth 24), and the share of layers with a sink from 0.53 to 0.22 and from 0.65 to 0.07 (lower in 2/3 seeds at depth 12, with 1 unchanged, and 3/3 at depth 24). Test accuracy falls from 55.9% to 53.3% and from 55.8% to 54.1% (lower in 3/3 seeds at depth 12 and 3/3 at depth 24). Massive activations do not fall consistently: at depth 12 they rise in 3/3 seeds (34.8 to 54.0), and at depth 24 the means barely move (60.9 to 61.3). This sigmoid is not the $-\log N$ rule of Section 6.

Change	Model	n	Sink readout	Massive activation	Quality
Sigmoid, depth 12	CIFAR-100 ViT	3	$0.026 \to 0.002$ (3/3)	$34.8 \to 54.0$ (0/3)	$55.9\% \to 53.3\%$ (3/3)
Sigmoid, depth 24	CIFAR-100 ViT	3	$0.033 \to 0.001$ (3/3)	$60.9 \to 61.3$ (2/3)	$55.8\% \to 54.1\%$ (3/3)
Output gate	char-level LM	3	$0.091 \to 0.041$ (3/3)	$9.64 \to 9.29$ (1/3)	$5.80 \to 5.79$ (2/3)
Four registers	WikiText LM	5	$0.106 \to 0.077$ -0.029 [-0.041, -0.017]	$12.25 \to 11.31$ -0.94 [-2.07, 0.50]	$4.824 \to 4.831$ 0.008 [-0.031, 0.046]
Four registers	CIFAR-100 ViT	5	$0.0198 \to 0.0106$ -0.0092 [-0.0183, 0.0034]	$29.8 \to 120.3$ 90.5 [61.1, 132.7]	$56.77\% \to 56.32\%$ -0.45 [-0.71, -0.23]
Value-null register	byte-level LM	5	$0.0889 \to 0.0763$ -0.0126 [-0.0196, -0.0070]	$8.05 \to 8.05$ 0.00 [-0.41, 0.40]	$2.0231 \to 2.0201$ -0.0030 [-0.0064, -0.0002]
Anchor gate	76.7M decoder	2	first token: $0.0111 \to 0.0068$ (2/2)	—	$25.80 \to 26.85$ (0/2) at 2,048: $88.1 \to 59.8$ (2/2)
Observational	GPT-2 Medium	384	$\rho = -0.37$ [-0.46, -0.28]	—	—
Observational	Pythia-1.4B	384	$\rho = -0.45$ [-0.56, -0.34]	—	—
Observational	Llama-2-7B	1,024	$\rho = -0.50$ [-0.56, -0.45]	—	—

Table 3: Trained-model evidence on the sum-to-one constraint. Each entry shows the control mean $\to$ the mean after the change; n counts seeds, or heads for the observational rows. With five seeds the paired change and its 95% interval follow; with two or three seeds, the number of seeds in which the readout is lower after the change, and every seed's value is in Table 12. No p-values. The sink readout is content-sink mass, except first-token mass for the anchor gate and, for the observational rows, ρ below. Quality is test accuracy for the ViTs, bits per byte for the value-null register, and perplexity otherwise; the anchor gate's second entry is its perplexity at 2,048 tokens, four times its training length. The sigmoid is unnormalized, with its output scaled by $1/N$. Content-sink mass in the causal models (the output gate, the register language model and the value-null register) rests on the value-norm test and a degenerate loading (see text). The observational rows measure public models as they are: ρ is the Spearman correlation between a head's total sink mass and its share of the summed per-head output norms (before the output projection), and its intervals, from resampling heads of one model, are descriptive. The anchor model's legacy LAMBADA scorer does not measure accuracy and is omitted. Production-ViT evidence, observational and with its own sink definitions, is in Table 4.

A headwise output gate (Qiu et al., 2025) is a diagonal gate on the softmax (Section 4), so it too lets a head attend less. In a hybrid causal character-level model with two softmax layers it more than halves the mean content-sink mass, from 0.091 to 0.041 (lower in 3/3 seeds, by 0.060, 0.047, 0.042), while first-token mass (0.038 to 0.040), massive activations (9.64 to 9.29) and perplexity (5.80 to 5.79) barely move.

7.2 Registers, an anchor gate, and what sink mass tracks

Extra tokens remove less. Adding four registers (Darcet et al., 2024) to a causal character-level WikiText language model lowers content-sink mass from 0.106 to 0.077 (change -0.029, 95% interval $[-0.041, -0.017]$, five seeds) with no detectable change in its per-character perplexity (4.824 to 4.831; $[-0.031, 0.046]$). Its massive activations go from 12.25 to 11.31, but the interval $[-2.07, 0.50]$ spans zero, so we claim no reduction.

In a CIFAR-100 ViT the same change raises massive activations in every seed, from 29.8 to 120.3 (change 90.5, $[61.1, 132.7]$), changes accuracy by -0.45 points ($[-0.71, -0.23]$), and leaves a sink-mass interval that spans zero ($[-0.0183, 0.0034]$).

A register with a null value would be a query-dependent null logit (Proposition 2). We trained one whose embedding is certified to be annihilated by every value map. In a causal byte-level model it cuts content-sink mass by only 14% (0.089 to 0.076; $[-0.020, -0.007]$, five seeds), against the 40% its prespecified test asked for, and not detectably more than a plain register token (certified minus plain -0.003, $[-0.010, 0.006]$). The register receives 0.054 of the attention, its own row included; bits per byte improve slightly ($[-0.0064, -0.0002]$), and massive activations show no detectable change. Its dial is a learned scalar per layer, added to every query's attention logit for the register. Setting the dial moves the register's mass monotonically (Spearman 0.987 across six dial settings in five seeds), but the median per-layer slope of the register mass's log-odds against

the dial is 0.38 ($R^2 = 0.95$), below the prespecified band of 0.7 to 1.3. That mass includes the register's own row, a floor of 1/129, which flattens the slope at low settings.

An anchor trades quality for length. A per-head anchor gate mixes each head's output with the token's own value, $o = (1 - a)PV + aV$, so its rows still sum to one. In a 76.7M-parameter RoPE decoder trained at 512 tokens it cuts perplexity at 2,048 tokens from 88.1 to 59.8 (lower in 2/2 seeds), but raises it at the training length from 25.80 to 26.85 (higher in 2/2 seeds), and it lowers the first token's mass from 0.011 to 0.007.

Sink mass and content energy. Observationally, heads that put more mass on sinks carry less of their layer's output. Across the heads of GPT-2 Medium, Pythia-1.4B and Llama-2-7B, a head's total sink mass and its share of the summed per-head output norms (before the output projection) have Spearman correlations -0.37, -0.45 and -0.50 over 384, 384 and 1,024 heads. The sink mass is the causal detector's, so it rests on the value-norm test and a degenerate loading, and intervals from resampling heads are descriptive only, since heads of one model are not independent. A prespecified within-head test asked about a sharper link: whether a head's sink mass tracks its contraction defect beyond the mixing term. In this test, sink mass is the mean first-token weight over the head's content queries, with first-token values idealized as null because their value contribution is unavailable. The result was negative: 0 of 7 public causal language models show a positive association, with decision statistics from -0.175 to -0.083 (Table 13).

Sink mass as an eviction rule. A prespecified test turned absorbed mass into a cache policy: on Pythia-1.4B, over coherent 2,048-token PG19 windows, it ranked cached tokens by how much of their content had drained into the sink through the layers and evicted the most absorbed first, always keeping the first token, the sink. Its retention certificate, a worst-case bound on how far the evictions can move the final-layer states, held with no violations and was non-vacuous on every window: it stayed inside the trivial bound. As a policy it lost. The policy comparison uses a linear model of the forward pass built from each layer's captured, head-averaged attention. At a 12.5% cache, the policy's mean relative deviation from the full cache's final-layer states exceeded that of H2O (Zhang et al., 2023), TOVA (Oren et al., 2024) and StreamingLLM (Xiao et al., 2024) by 0.018, 0.016 and 0.014. Matching H2O's deviation at a 14.0% cache took a 37.5% cache.

What the records show. Two ways to attend less, sigmoid attention and an output gate, remove much of the sink mass; sigmoid pays for it in accuracy. Extra tokens remove less. Registers lower sink mass in a language model but raise a ViT's massive activations, and a register whose embedding is certified value-null removes only 14%, not detectably more than a plain one. These are small models with few seeds and one change each, and none of them tests the two-sided law. The trained harness was built to, but at 125M no retrieval head reached the registered identification threshold (Section 8).

7.3 Learned nulls and positions in byte-level models

Learned nulls across training lengths. In the tested byte-level language models with no positional encoding and a learned null, training does not hold abstention fixed as the context grows. We trained 25 models of 594,320 parameters for 1,500 steps at 5 lengths from 64 to 1,024 tokens, 5 seeds each. The learned null, relative to the mean content logit, does not rise with the training length (slope -0.14 per e-fold, 95% interval $[-0.18, -0.10]$), and the log-odds of abstaining fall at -0.93 per e-fold of the training length ($[-0.98, -0.88]$), close to the fixed-null law's -1. Within each model, at fixed weights, they fall at -1.04 to -1.01 per e-fold of the context in the models trained at 128 tokens or more (-1.31 at 64 tokens), as medians over the heads of the five seeds trained at each length. The measured within-model slopes are close to the stationary law's prediction in Section 5, except in models trained at the shortest length. These are small models, and the result concerns abstention, not retrieval.

Positions and nulls in small byte LMs. The same byte-level model, trained at 256 bytes with three positional schemes (none, rotary, rotary on a quarter of each head) and four null rules (none, a learned null, a learned null plus $\ln n$, an output gate), 5 seeds each, separates the positional channel from the null. Without positional encoding the background grows like n under every rule (γ 1.08 to 1.13 inside the training

length), and a fixed null's log-odds fall by the stationary law (slope −1.08, 95% interval [−1.09, −1.08]). Adding $\ln(n/N_{\text{train}})$ to the null at evaluation then holds its odds level beyond the training length (slope −0.02, against −1.02 without it) and moves the loss by less than 10^{-4} bits per byte.

With rotary positions the background grows like $n^{0.12}$ to $n^{0.19}$ with no null, an output gate or a fixed null alike, so a fixed null's slow decay is the kernel's default. Only a null that itself adds $\ln n$ raises the background growth, to $n^{0.97}$ (difference 0.78, [0.68, 0.89]), and those heads learn less local attention: the median head's top key is among the last 8 in 46% of rows, against 82% with a fixed null.

Beyond the training length, without positional interpolation, the heads that held content furthest below a fixed null raise their content logits most (Spearman 0.87, [0.78, 0.92]), and a fixed null's length penalty exceeds plain softmax's by 0.12 and the output gate's by 0.11 bits per byte (5 and 4 of 5 seeds). Whether rotating only a quarter of each head reduces this length penalty is undetermined: it did so in 2 of 5 seeds. With rotary positions, adding YaRN's temperature to interpolation raised these models' length penalty in every checkpoint (Table 7). These are small models on one corpus, at one training length. See Figure 4.

7.4 Production vision transformers

Most of the measurements above come from small models trained for the purpose. An observational census of seven pretrained ViTs gives a check at production scale: DeiT-B, DeiT3-B, ViT-B pretrained on ImageNet-1k, ViT-B with its original ImageNet-21k pretraining, and ViT-S, ViT-B and ViT-L with AugReg ImageNet-21k pretraining, each measured on the same 1000 ImageNet validation images and re-analysed here from the census records (Tables 4 and 5, Figure 5). The census has two sink definitions, which we keep apart. A CLS-readout sink (sink_cls) is a patch token in the top quartile of the CLS token's attention and the bottom quartile of composed value norm. A π-outlier sink (sink_H) is a patch token whose stationary mass under the head-averaged patch chain is an outlier while its value norm and cluster loading are low. A bidirectional ViT has no causally absorbing token, yet 6 of the 7 checkpoints grow a sink_H population, born between layers 6 and 11 with peak median counts of 3–10 tokens out of 196. Birth is the first layer at which at least half the images carry a flagged token. DeiT3-B grows none; instead its CLS token receives 11–36% of head-averaged attention across layers. Where a population exists, 87–99% of its tokens lie inside the CLS-readout corner.

How long the sink_H set persists across depth differs between checkpoints (Figure 5a). The set locks at the first layer whose pooled Jaccard overlap with the next layer reaches 0.7, and dissolves at the first layer after its peak whose overlap with the layer before falls below one half. Its largest adjacent-layer Jaccard overlap is 0.074 in DeiT3-B, which has no population and few flagged tokens, and 0.97 in ViT-L; ViT-L locks, turns over across layers 17–19, locks again and dissolves at layer 22. Neither a set that re-forms at every layer nor a fixed register describes these models, and with one checkpoint per setting the difference between them cannot be attributed to the training recipe, the data or the scale. The census registered its predictions only for ViT-B with its original ImageNet-21k pretraining (Appendix D.2); the one about persistence, that the set locks and then dissolves, misses the lock rule at its point estimate ($J_{\max} = 0.688$ against 0.7), although its interval spans the threshold.

These in-sequence sinks are not value-null. Zeroing a sink token's value at one layer, with attention unchanged, moves the mean-centred output logits 1.5–1.9 times as much as zeroing a typical patch for sink_cls, and 2.4–10.6 times as much for sink_H, in all 6 checkpoints with ground truth; removing the key adds a renormalization effect on top. sink_H delivers latest in depth (transport centre of mass at 0.73–0.86 of depth), partly because the set exists only from its birth layer on. Both populations carry content of their own as well as holding attention that would otherwise go to other keys, so the value-null factorization of Section 4 does not describe them. The census detector differs from the one behind Table 3, so its counts are not compared with that table's.

8 A trained harness: which rule keeps both sides in vision transformers and 125M language models

Section 6 trained single heads on fixed features. The harness trains whole vision transformers and language models and changes one thing per model, the rule by which its heads normalize attention, while the backbone,

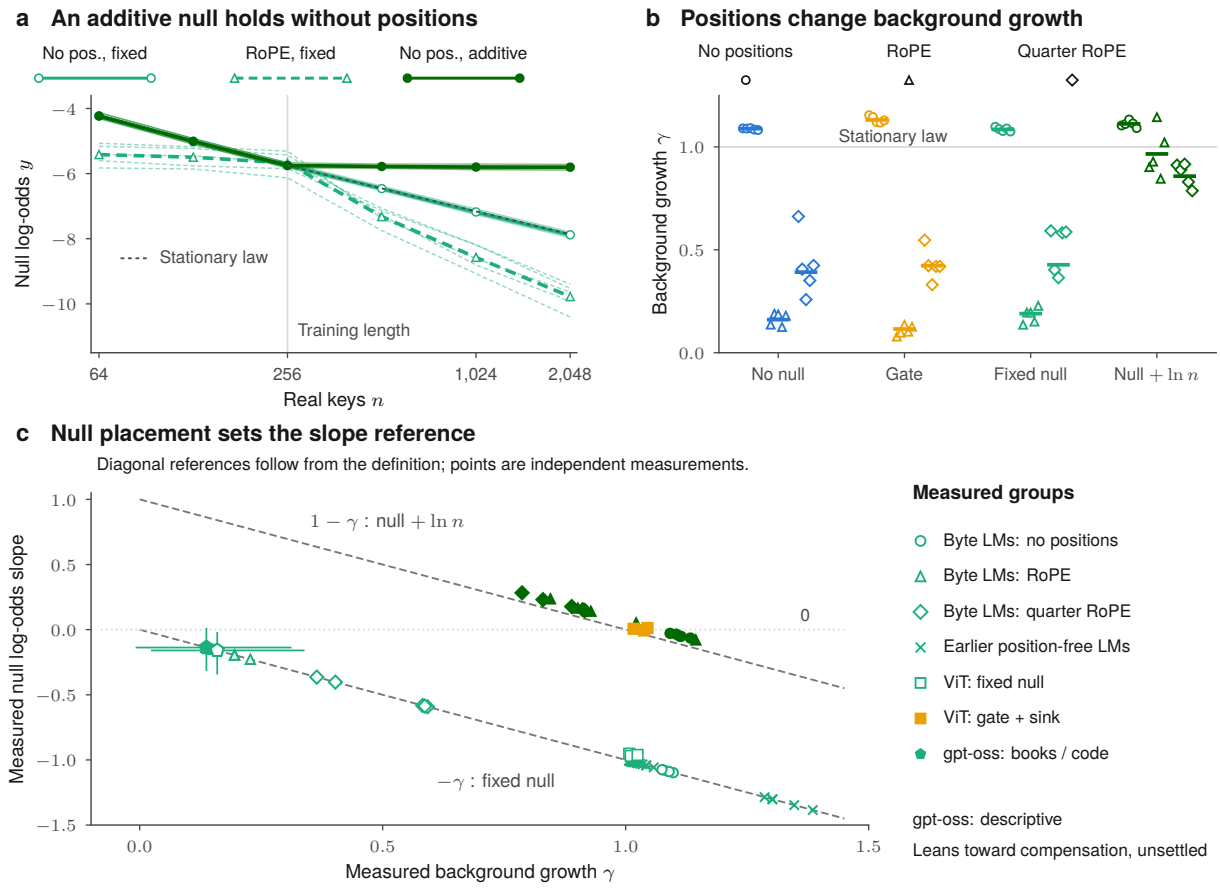


Figure 4: Positions change the background's growth, and null placement sets how that growth affects abstention. (a,b) 0.6M-parameter byte language models, 5 seeds per arm, trained at 256 tokens on one corpus. (a) Measured null log-odds $y = \log[p_\varnothing/(1-p_\varnothing)]$: thin curves are seed medians over heads, bold curves their mean. The additive null changes the fixed null by $\ln(n/N_{\text{train}})$ only beyond training, with no positions; the vertical rule marks the training length. The dashed stationary-law reference has slope -1.02, measured in the earlier position-free models trained at the same length, and is anchored to the fixed-null mean at training. (b) Background log-sum slope γ over 64–256 keys: each dot is a seed's median head slope; the short ticks are medians over all heads and seeds, as in the assessment. Shapes identify positions. Colors retain the null rule across panels. (c) Both coordinates are measured independently over matching key counts. The two diagonal references follow from the definition of log-odds for a fixed null logit ($-\gamma$) and a null logit growing as $\ln n$ ($1-\gamma$); the horizontal reference is zero slope. They are not fitted trends or a theorem that gates are constant. Byte-LM circles/triangles/diamonds are the 5 seeds of each positional null arm over 64–256 keys (fixed null open, $+\ln n$ filled). Crosses are all 25 earlier position-free models (training lengths 64–1,024); each fits the saved counts $\{128, 512, 2048, N_{\text{train}}\}$, removing repeats, then takes the median head slope. Squares (fixed null open, gated filled) are 6 checkpoints (3 seeds per rule) of one ViT-S/16 size, untempered CLS rows over 224–896 px, beyond the 224 px training resolution; they use recorded medians over images, heads and all 12 layers on the first 1,000 images. For the ViT, the background sums patches only, while null odds include CLS among the real keys; the fixed-null points therefore need not lie on $-\gamma$. The two large pentagons are gpt-oss books (filled) and code (open) medians with marginal interquartile ranges over the same 700 sink-using dense heads, of the slope between query positions 256 and 4,096 in natural documents, which is shallower than the least-squares medians over every position of that range (Table 6): descriptive; the prespecified test on these heads was undecided by its rule. The byte-model contrasts are prespecified; the cross-model synthesis and ViT null slopes are descriptive. Causal interpretation is limited to the controlled positional contrasts.

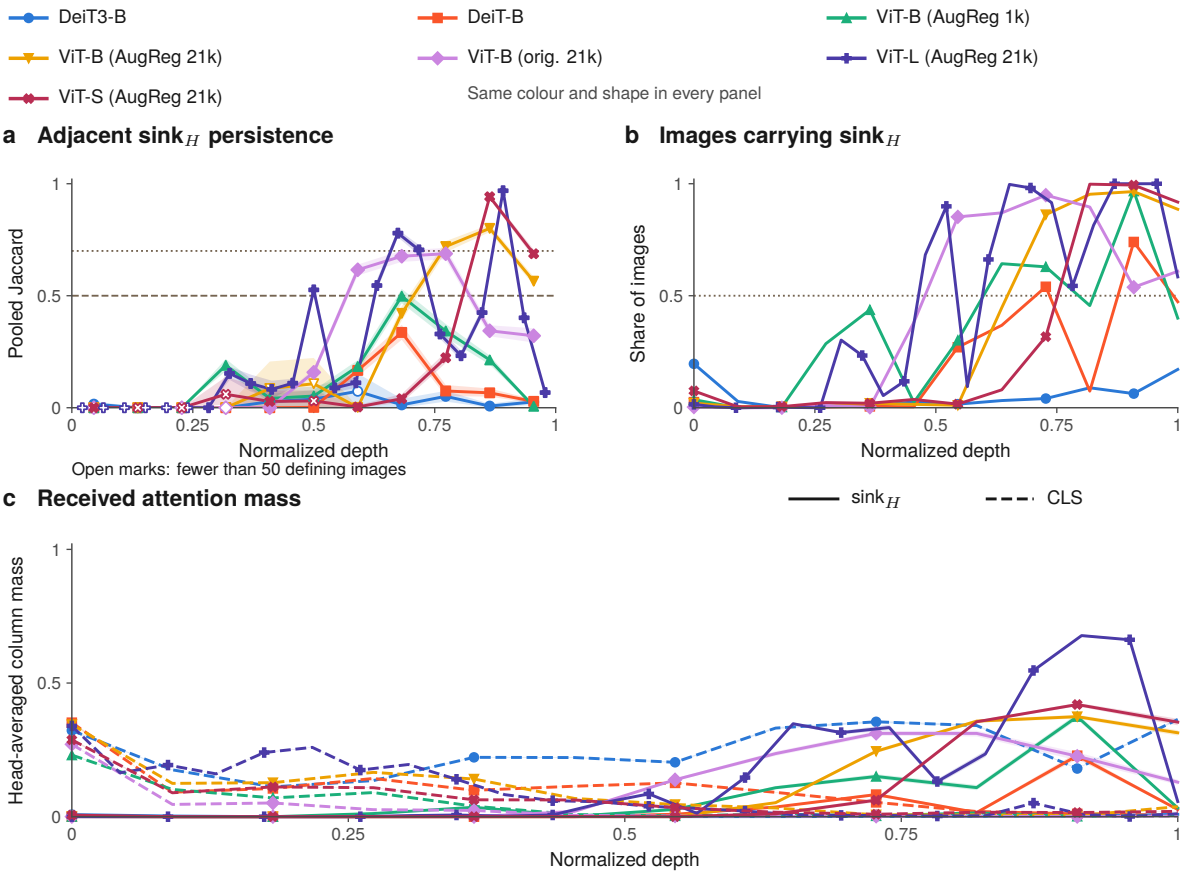


Figure 5: Sink persistence differs across the seven pretrained ViT checkpoints. (a) Pooled adjacent-layer Jaccard of the sink_H set, with hollow model markers where fewer than 50 images define the pair; dotted and dashed rules mark 0.7 and 0.5. (b) Share of images carrying sink_H; the dotted rule marks 0.5. (c) Head-averaged received attention mass: solid sink_H, dashed CLS. Checkpoint colours and marker shapes are fixed across panels. Bands show the recorded 95% cluster-bootstrap intervals over images: 400 resamples for Jaccard and 2,000 for census mass. 1,000 images per checkpoint. Patch tokens only for sink_H (the pi-outlier set); sink_cls is the CLS-readout corner, whereas CLS mass is the CLS column itself. The lifecycle prediction is registered for ViT-B original-21k only; other checkpoints are robustness evidence. One checkpoint per setting cannot separate recipe, data and scale. The detector differs from the one behind Table 3, and their counts are not compared. Source: the census records (Table 16).

Lifecycle of the sink_H set

Checkpoint	Blocks	Birth	Peak (median)	Largest image share	Lock	Dissolution	Absent before L5	Lock then dissolve	Tier
DeiT3-B	12	—	—	0.196 (L0)	—	—	FALSIFIED	FALSIFIED	robustness
DeiT-B	12	8	L10: 3	0.740 (L10)	—	11	SUPPORTED	FALSIFIED	robustness
ViT-B (AugReg 1k)	12	7	L7: 3	0.967 (L10)	—	8	FALSIFIED	FALSIFIED	robustness
ViT-B (AugReg 21k)	12	8	L8: 4	0.965 (L10)	8	—	SUPPORTED	FALSIFIED	robustness
ViT-B (orig. 21k)	12	6	L8: 9	0.949 (L8)	—	10	SUPPORTED	FALSIFIED	registered
ViT-L (AugReg 21k)	24	11	L20: 10	1.000 (L20)	15	22	SUPPORTED	SUPPORTED	robustness
ViT-S (AugReg 21k)	12	9	L9: 5	0.998 (L9)	9	—	FALSIFIED	FALSIFIED	robustness

Persistence and definition overlap

Checkpoint	J max (pair)	95% CI	Images	H inside C	C that is H
DeiT3-B	0.074 (L6→7)	[0.000, 0.153]	40	0.519	0.009
DeiT-B	0.336 (L7→8)	[0.303, 0.370]	643	0.989	0.063
ViT-B (AugReg 1k)	0.498 (L7→8)	[0.467, 0.531]	807	0.872	0.073
ViT-B (AugReg 21k)	0.801 (L9→10)	[0.784, 0.818]	987	0.966	0.113
ViT-B (orig. 21k)	0.688 (L8→9)	[0.663, 0.714]	987	0.971	0.253
ViT-L (AugReg 21k)	0.969 (L20→21)	[0.963, 0.974]	1000	0.891	0.212
ViT-S (AugReg 21k)	0.942 (L9→10)	[0.931, 0.952]	999	0.956	0.086

Table 4: Per-checkpoint existence, persistence and definition overlap. Zero-median trajectories have no peak or inferred dissolution. Verdicts FALSIFIED/SUPPORTED refer to the registered (ViT-B original-21k) or robustness (other checkpoints) lifecycle predictions. 1,000 images per checkpoint. Head-averaged; patch tokens only (CLS column mass is labelled separately). sink_H is the pi-outlier set; sink_cls is the CLS-readout corner. Source intervals: 400-resample cluster bootstrap over images; census means: 2,000 resamples. Lifecycle predictions are registered for ViT-B original-21k only, panel entries are robustness; channel logit space is registered and feature space is robustness. The detector differs from the one behind Table 3, and their counts are not compared.

the optimizer, the data and the evaluation stay fixed (Appendix D). Its language-model leg trains 11 rules in 125M-parameter decoders: softmax; softmax with four value-null registers; off-by-one softmax; a learned sink logit; the sink plus $\log n$; sigmoid attention with a $-\log n$ bias and a norm on its output; a headwise output gate; a signed gate, which mixes the softmax row with a softmax of the scores' magnitudes that carries their signs; the gate with a learned sink; the learned sink with a gate on each value, as in concurrent work (Wang, 2026); and a key check like that of Appendix B, a head that reads a maximum: it routes over its eight largest scores behind a gate on the best one. A seed fixes the token stream, so comparisons between rules are paired by seed. The learned sink and the gate with a learned sink are also tempered at evaluation, by the schedule of Theorem 8, with N the number of keys each query sees and the factor $\max(1, \ln N / \ln N_{\text{train}})$, so that no score is softened. The language models are evaluated to eight times their training length, natively and with YaRN (Peng et al., 2024), the registered primary positional treatment, by three measurements: the perplexity ratio D on long books; passkey retrieval; and a probe whose query names a fact that is either present in the context or absent from it. The probe is read on retrieval heads, chosen on separate identification trials by their attention to the matching fact.

8.1 Vision transformers: a first result and its confirmation

The ViT subset gives the first trained result. The harness's ViT subset trained ViT-S/16 on ImageNet-100 at 224 px with five attention rules (softmax, value-null registers, a learned sink, an output gate, and the gate with a learned sink), three seeds each, and evaluated every model at 224, 320, 448 and 640 px, which multiplies the number of patch tokens by 1, 2.0, 4.0 and 8.2. Unless stated otherwise, every resolution above 224 px in this section is the 224 px crop upsampled bicubically, with the position embeddings interpolated bicubically, so it adds tokens but no image content. Its one registered prediction failed. It held that the output gate, alone or with a tempered sink, lowers both the content-sink mass (the attention absorbed by patch tokens that act as sinks) and the ratio of maximum to median absolute activation below softmax's, in every seed at 224 px. Content-sink mass fell in all three seeds, but the activation ratio rose in one seed for

One-layer counterfactual ratios against a typical patch: ratio (δ) verdict

Checkpoint	sink_H transport	sink_H modulation	sink_cls transport	sink_cls modulation
DeiT3-B	3.84 (0.60) F	6.46 (0.66) S	1.54 (0.41) F	1.57 (0.46) S
DeiT-B	5.17 (0.84) F	10.05 (0.92) S	1.60 (0.50) F	1.60 (0.52) S
ViT-B (AugReg 1k)	5.16 (0.94) F	13.34 (0.97) S	1.88 (0.56) F	2.28 (0.69) S
ViT-B (AugReg 21k)	10.65 (0.90) F	29.46 (0.99) S	1.74 (0.58) F	1.86 (0.62) S
ViT-B (orig. 21k)	2.35 (0.56) F	6.29 (0.85) S	1.65 (0.47) F	2.14 (0.60) S
ViT-L (AugReg 21k)	4.63 (0.91) F	8.42 (0.96) S	1.62 (0.53) F	1.86 (0.61) S
ViT-S (AugReg 21k)	—	—	—	—

Delivery depth and attention mass

Checkpoint	sink_H depth COM	sink_cls depth COM	Luminous depth COM	Chain5	Peak sink_H attention mass	95% CI	CLS mass, second half
DeiT3-B	0.727	0.598	0.537	True	0.009 (L11)	[0.007, 0.011]	0.296
DeiT-B	0.810	0.697	0.558	True	0.228 (L10)	[0.216, 0.240]	0.051
ViT-B (AugReg 1k)	0.748	0.541	0.576	False	0.374 (L10)	[0.364, 0.386]	0.006
ViT-B (AugReg 21k)	0.862	0.627	0.600	True	0.374 (L10)	[0.367, 0.381]	0.022
ViT-B (orig. 21k)	0.751	0.523	0.463	True	0.312 (L8)	[0.303, 0.321]	0.002
ViT-L (AugReg 21k)	0.797	0.615	0.544	True	0.677 (L21)	[0.673, 0.682]	0.015
ViT-S (AugReg 21k)	—	—	—	—	0.420 (L10)	[0.413, 0.426]	0.019

Table 5: Ratios compare one-layer counterfactual output changes with a typical patch; F and S mark the two registered channel predictions as falsified or supported: transport near zero (falsified in every checkpoint with ground truth) and modulation large (supported in every one). δ is Cliff's delta of those changes. All ratios are in logit space, the registered space. DeiT3-B sink_H ratios use only two layers; ViT-S has no ground truth. COM is the centre of mass as a fraction of depth; Chain5 marks whether the five populations' centres are ordered background < mid < luminous < sink_cls < sink_H. Census mass is a size statistic partly induced by selection. 1,000 images per checkpoint. Head-averaged; patch tokens only (CLS column mass is labelled separately). sink_H is the pi-outlier set; sink_cls is the CLS-readout corner. Source intervals: 400-resample cluster bootstrap over images; census means: 2,000 resamples. Lifecycle predictions are registered for ViT-B original-21k only, panel entries are robustness; channel logit space is registered and feature space is robustness. The detector differs from the one behind Table 3, and their counts are not compared.

both variants. That ratio ranges from 25 to 31 across softmax's own seeds, the outliers it measures are mild in these models, and W8A8 quantization moves any variant's seed-mean accuracy by at most 0.2 points.

The registration reports the other resolutions without a pass-or-fail criterion, so what follows is descriptive. Tempering the trained sink at evaluation, multiplying every score, the sink's included, by $\log N / \log N_{\text{train}}$ (the schedule of Theorem 8), leaves accuracy at 224 px unchanged by construction. It raises accuracy by 0.9, 2.0 and 3.5 points at 2.0, 4.0 and 8.2 times the training token count, by 3.4 points for the gated sink at 8.2 times, and by at least 3.1 points in every seed at 8.2 times. At 640 px the tempered sink beats untempered softmax by 4.0 points (95% interval [3.1, 5.0]), while the untempered sink, the output gate and the value-null registers each stay within 0.6 points of softmax. Content-sink mass falls with resolution under softmax, from 0.042 to 0.037, and stays between 0.041 and 0.043 when tempered (Figure 6), consistent with a share that dilutes as tokens are added and a log-N schedule that holds it. That share is the patch tokens' sink mass, not the learned null's. The records store no per-head null logit, but a head-averaged proxy for the null's mass falls with slope -1.02 in $\log N$ untempered, near the stationary law's -1, and still -0.83 tempered, so the schedule does not hold the null. These are three seeds of one model size on one dataset. The language-model harness registered the corresponding prediction, that the tempered arms degrade least with length, and it failed (below).

A prespecified confirmation. A second evaluation of the same fifteen checkpoints was specified and frozen before any of its numbers existed. It added two measurements: accuracy at 896 px, 16 times the training token count and a resolution not evaluated before; and softmax tempered the same way, its scores multiplied by $\log N / \log N_{\text{train}}$ with no sink. Both predictions about the extrapolation held. At 896 px, tempering raised accuracy by at least 3.7 points in every seed of the learned sink and the gated sink, against a prespecified threshold of 2.0 points (seed means 4.9 and 5.0). The seed-mean gain was also larger than at

640 px for both, by 1.4 and 1.6 points, each with a 95% interval above zero. The attribution test, fixed in advance at 640 px, came out in its branch for log-N scaling alone. There, tempering the sink gained 3.48 points and tempering plain softmax 4.16, so the sink's gain minus softmax's is -0.68 points (95% interval $[-1.30, -0.04]$). At 896 px the two gains were 4.87 and 4.98 points (difference -0.11, 95% interval $[-1.26, 0.92]$). The accuracy gain therefore comes from scaling the scores with $\log N$, which the sink's tempering includes, and the sink is not needed for it; this is not a claim that sinks have no other use. At 896 px every untempered rule's seed-mean top-1 accuracy lies between 57.2 and 58.3 percent, and the three tempered variants reach 62.8 to 63.2 percent. The test reuses the subset's checkpoints and validation images, so it adds measurements, not seeds or images. It is prespecified, not publicly registered; so is every other registration in this section, each frozen before its run.

8.2 Attention rows at higher resolution, on two input routes

A registered re-evaluation of nine of these checkpoints (softmax, the learned sink and the gate with a learned sink, three seeds each) summarized the CLS rows' attention at 224 to 896 px, untempered and tempered, on two input routes: the upsampled crops above, and the original images resized to each resolution. On the upsampled crops the background grows as dilution predicts. From 224 to 896 px the log-sum of the CLS rows' content logits rises with slope 1.01 to 1.04 in $\ln N$, against 1 for the stationary law, while the top logit rises far more slowly (slope 0.25 to 0.30) and the rows' entropy rises at the dilution's rate (slope 0.96 to 0.98), in all 9 checkpoints.

Tempering does not restore the trained attention. In the last four layers, where regions, entropy and persistence are compared, it moves the rows' attention further from its 224 px form, region by region (total variation up by 0.027 to 0.045 in 9 of 9 checkpoints, from 0.33 to 0.37 untempered), and undoes only 17% to 26% of the entropy's rise. The region a row preferred at 224 px, or one of its eight neighbours, keeps its top key in only 0.20 to 0.23 of rows at 896 px (chance 0.046), and the per-image gain does not track that persistence (Spearman -0.040, -0.023 and $+0.028$ for the three rules; the softmax's 95% interval lies below zero). Why tempering helps these models is therefore open.

The gain itself holds on resized originals: 4.87, 4.65 and 4.54 points at 896 px (95% intervals [4.29, 5.46], [3.94, 5.41] and [3.93, 5.14]), against 4.99, 4.87 and 4.99 on the upsampled crops in the same re-evaluation. Most originals are themselves upscaled at that size: their median shorter side is 375 px, and 95% of them carry less than half the detail of a native 896 px image.

Per head, the learned sink's log-odds fall at -0.96 per unit of $\ln N$ (-0.97 to -0.95), near the stationary law's -1, while with the output gate added to the sink abstention holds level (slope -0.006 to $+0.013$): a null in the scores competes with a background that grows like N, and a gate does not. The re-evaluation was registered before it ran. It reuses the subset's three seeds per rule of one model size, and its row summaries cover the first 1,000 validation images. See Figure 7.

8.3 Language models at 125M

The harness's language-model leg trained a 125M-parameter decoder with each of 11 attention rules, 3 seeds each, at 1,024 tokens, and evaluated it to 8,192 tokens, with YaRN as the registered primary positional treatment. Its lead prediction failed (Figure 8a–c). On the same checkpoints, tempering raised long-context degradation against softmax. The perplexity ratio D divides a checkpoint's long-context perplexity by its perplexity on the same books' training-length prefixes. At 8 times the training length, this ratio exceeded softmax's by $+0.27$ for the tempered learned sink (95% interval $[+0.21, +0.34]$) and by $+0.18$ for the tempered gated sink ($[+0.14, +0.22]$). Of that excess, tempering itself, against the same checkpoint untempered, added 0.10 to 0.17 in each seed. Tempering also did not hold abstention within the registered bound. None of the probe's retrieval heads met the separated margins that Theorem 8 needs, so its guarantee was not exercised.

Fixed sinks behaved differently on text and on the abstention probe. On text, the sink log-odds of off-by-one softmax and of a learned sink fell with slopes -0.14 and -0.12 in $\log(i+1)$ inside the training length (-0.07 for the sink plus $\log n$), and -0.29 and -0.28 beyond it with YaRN, against -1 for a stationary background, and as shallow as gpt-oss's dense layers (-0.20 and -0.19, Section 9); natively, beyond the training length,

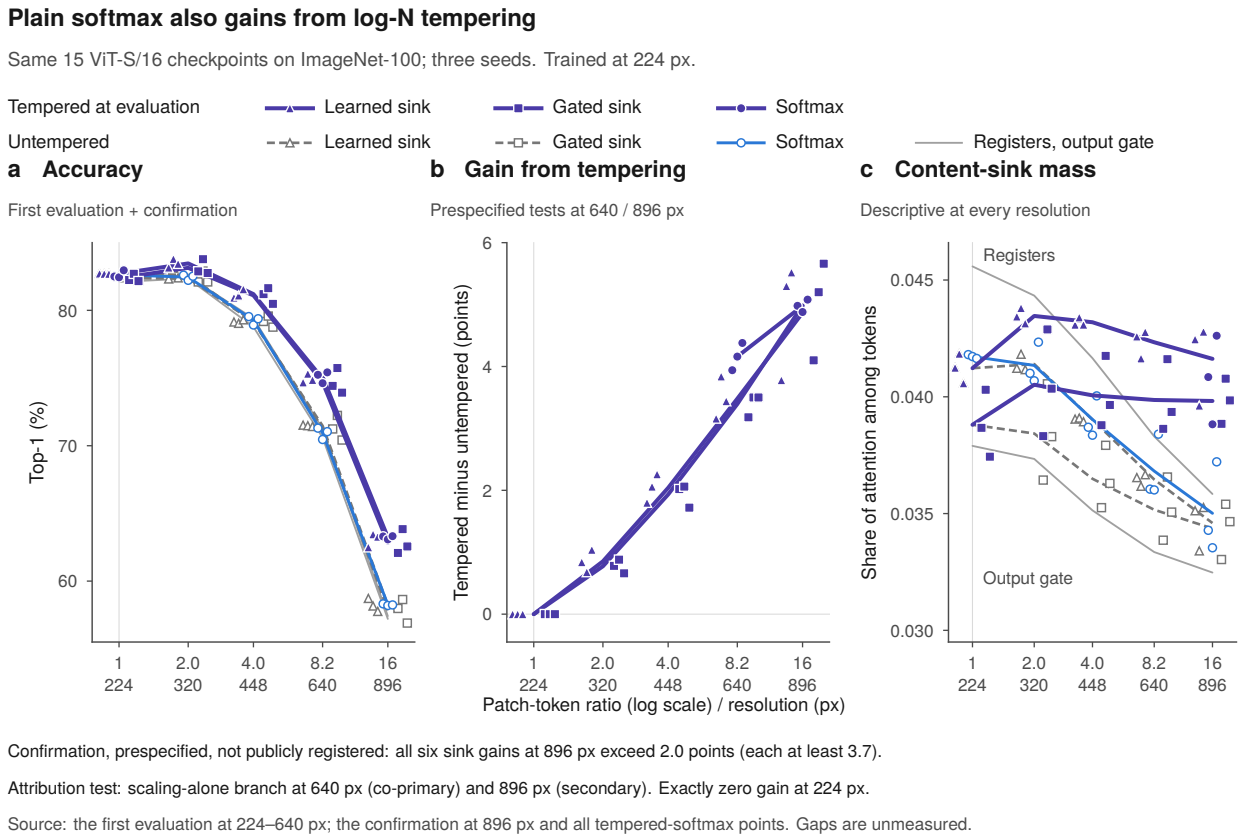


Figure 6: Log-N evaluation tempering improves accuracy through 16 times the training patch-token count, and plain softmax gains at least as much as the learned sink. The same fifteen ViT-S/16 checkpoints were trained on ImageNet-100 at 224 px (five arms, three seeds). (a) EMA top-1 accuracy on 5,000 validation images. (b) Tempered minus untempered accuracy of the same checkpoint, paired by seed and image, for the learned sink, the gated sink and plain softmax. The confirmation's extrapolation test holds: all six sink and gated-sink gains at 896 px are at least 3.7 points, above the 2.0-point threshold (means 4.9 and 5.0). Its attribution test takes the scaling-alone branch: softmax versus sink gains are 4.16 versus 3.48 points at 640 px (co-primary) and 4.98 versus 4.87 at 896 px (secondary). (c) Descriptive content-sink mass, averaged over 12 layers on 64 fixed validation images: attention absorbed by patch-token sinks after renormalizing among real tokens, excluding a sink's or gate's abstention; CLS is not content. Lines join three-seed means at true patch-token ratios $(r/224)^2$; marks show seeds, displaced horizontally by checkpoint and seed. Registers and the output gate show means only. Filled marks are tempered; open marks are untempered. Tempering multiplies scores, including any sink, by $\log N / \log N_{\text{train}}$, counting patches and CLS in N, hence is exactly the identity at 224 px. The first evaluation supplies 224, 320, 448 and 640 px; the confirmation supplies 896 px and every tempered-softmax point. For tempered softmax, accuracy exists only at 224/640/896 px and content-sink mass only at 896 px; missing points are neither imputed nor connected across. The first evaluation's resolution trends and panel (c) remain descriptive; the confirmation is prespecified, not publicly registered, and reuses the same checkpoints and validation images.

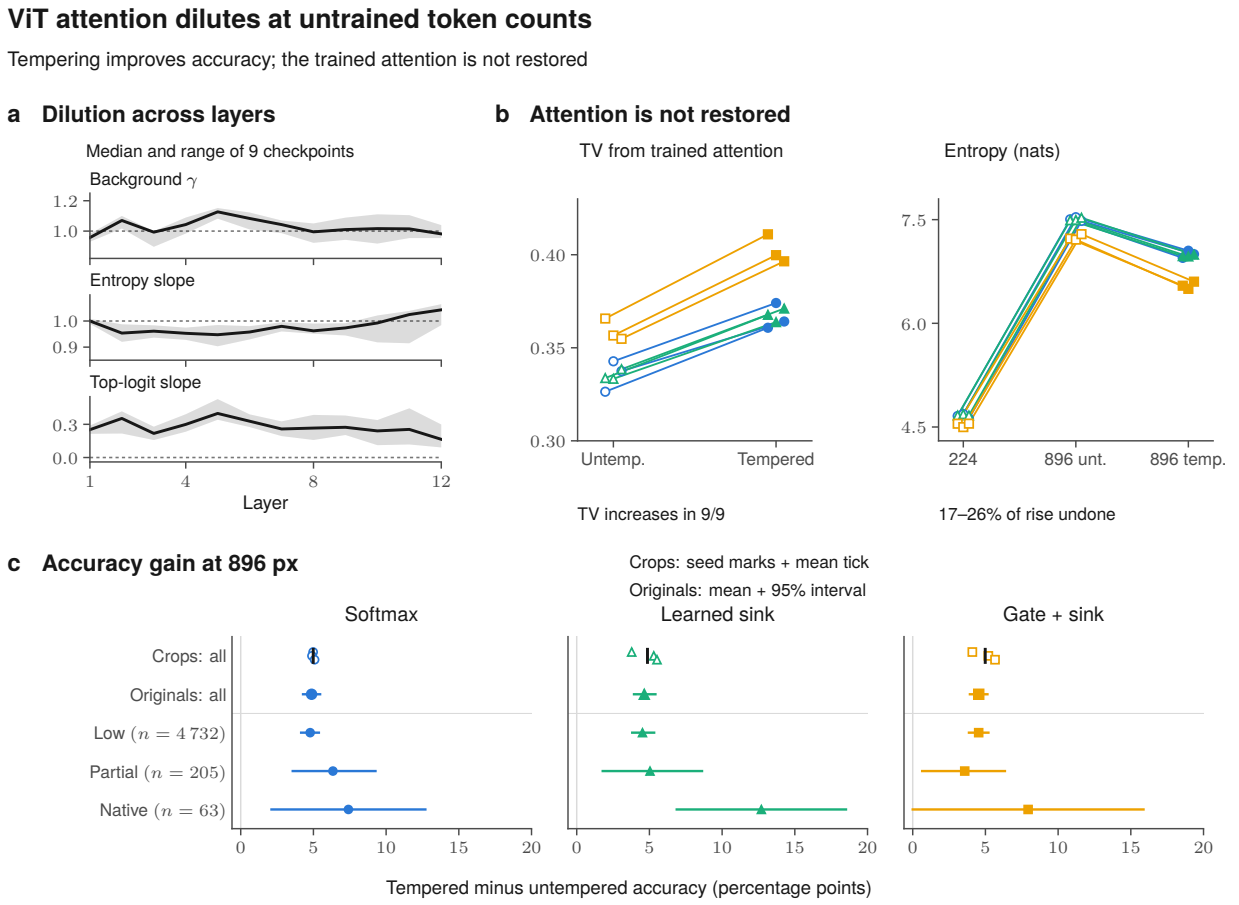


Figure 7: At untrained token counts, ViT attention dilutes, and log-N tempering raises accuracy without restoring the trained attention. A registered re-evaluation uses 9 checkpoints (3 rules × 3 seeds), one ViT-S/16 size on ImageNet-100. (a) Layer profiles of the registered untempered CLS slopes against $\ln N$ (patch count) over 224–896 px: median across checkpoints and full range; the background log-sum and entropy have stationary-law reference 1, the top logit reference 0. (b) Paired checkpoint summaries, CLS layers 9–12 on upsampled crops: the registered total-variation (TV) comparison with the 224 px regional attention, and descriptive entropy in nats. Open marks are untempered; filled marks are tempered. TV rises in 9 of 9; tempering undoes only 17–26% of the entropy rise (descriptive). (c) Tempered minus untempered top-1 accuracy at 896 px: the registered overall comparison on resized originals and descriptive crop/detail comparisons. For resized originals (O), bars are the recorded 95% paired hierarchical bootstrap intervals, resampling seeds then matched images. For upsampled crops (U), the three small marks are seed gains and the vertical tick their mean; no image-bootstrap interval is available in these summaries. The lower three rows give original-detail strata, with images per seed printed at left: low is below half, partial is at least half but below full, and native is full linear detail. These strata are descriptive. Accuracy uses 5,000 validation images; row summaries use the CLS rows of the first 1,000. The registered dilution and bounded-top/entropy tests held, the restoration and gain–persistence tests were falsified, and leader persistence was inconclusive, meeting neither its held nor falsified threshold. Why tempering helps remains open. Most resized originals are also upscaled at 896 px: the median shorter side is 375 px, and 95% carry less than half the native detail.

they fell at about −6.2. On the probe, with YaRN, they kept about a tenth of their null mass at 8 times the training length, from 0.028 to 0.003 and from 0.059 to 0.007. Here YaRN's temperature multiplies the content scores but not the sink logit (Appendix D), the premise violation of Section 5; applying it to the sink as well leaves the learned sink's null mass at eight times almost unchanged (0.0043 against 0.0034, three-seed means on the re-evaluation's probe), so the loss does not come from that asymmetry. The length-calibrated sink lost abstention under YaRN (a change of −5.2 in log-odds) and kept its match weight (−0.13), the reverse of its pilot pattern.

Every rule that gives attention a null removed the first-token sink (0.039 of the attention under softmax, about 0.003 or less with a null) and cut massive activations by 67% to 82%; a signed gate without a null did not (its 7% mean cut changes sign across the three paired seeds). Sigmoid aside, every rule with a null also lowered perplexity at the training length, by 1.2 to 2.5 percent; sigmoid attention raised it by 2.8 percent. With YaRN, softmax and the output gate degraded least with length (D of 1.22 and 1.23 at 8 times).

At this scale no retrieval head reached the registered identification threshold (largest weight 0.247 against 0.3), and no rule abstained behaviourally on the probe, so these abstention results describe attention weights, not behaviour. With YaRN, the head that reads a maximum 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 the matching key fell far down the ranking and the background maximum grew with length. Trained models at 350M parameters, and the cost of dropping sum-to-one at depth, are not yet measured.

8.4 Separating interpolation from temperature

YaRN combines position interpolation (Chen et al., 2023) with a log-length temperature. A prespecified re-evaluation of the same checkpoints (softmax, the learned sink, the output gate and the head that reads maxima; 12 checkpoints) applied YaRN's two parts one at a time. Interpolating the rotary frequencies alone carries 87% of YaRN's reduction in the log perplexity ratio at 8 times the training length (median; 70% for softmax, 99% for the head that reads maxima; Figure 8d). YaRN's temperature adds to it beyond the training length in 12 of 12 checkpoints, and alone it produces the in-range shift in the sink's log-odds (regressed on the native log-odds below the training length, the shift has slope 0.39 to 0.40, against −0.03 for interpolation). In these models sharpening helped only on interpolated positions. The schedule of Theorem 8 on interpolated positions lowers the loss beyond the training length in 9 of 9 checkpoints, by 0.20 to 0.50 nats per token, and raises softmax's passkey accuracy at 8 times from 0.65 to 0.83. Stacked on YaRN, which already tempers, it raises the loss in 9 of 9 (by 0.08 to 0.18 nats per token). YaRN's temperature alone, without interpolation, worsens long-context perplexity in 12 of 12 checkpoints. The harness's tempered arms were evaluated under YaRN, whose temperature also multiplied their sink logit, so every score carried one multiplier, as Theorem 8 assumes; their failure was over-sharpening. On these checkpoints YaRN's own temperature beats the schedule of Theorem 8 beyond the training length in 9 of 9. The passkey gain, the count for the temperature alone and this last comparison were not among the re-evaluation's prespecified predictions and are descriptive.

9 A prespecified test on gpt-oss

gpt-oss-20b (OpenAI, 2025) learns one sink logit b_h per head, so it carries the fixed-sink half of the law in a production model. At query position i a head's sink log-odds are exactly

$$y_h(i) \;=\; b_h - \mathrm{LSE}_{j\in K(i)}\, s_{ij} \;=\; b_h - \log n(i) - m_h(i),$$

where $K(i)$ is the set of $n(i)$ visible keys and $m_h(i)$ is their mean background level, the logarithm of the mean of $e^{s_{ij}}$. If the background is stationary, y_h falls against $\log(i+1)$ with slope -1; a network that lowers its background logits like $\log n$ holds it flat (Section 5). The model alternates 12 dense layers, where $n(i) = i+1$, with 12 banded layers whose window holds 128 keys, where $n(i) = \min(i+1, 128)$. The mechanism is the same and only the key count differs, so the banded layers are a built-in control: beyond the window they

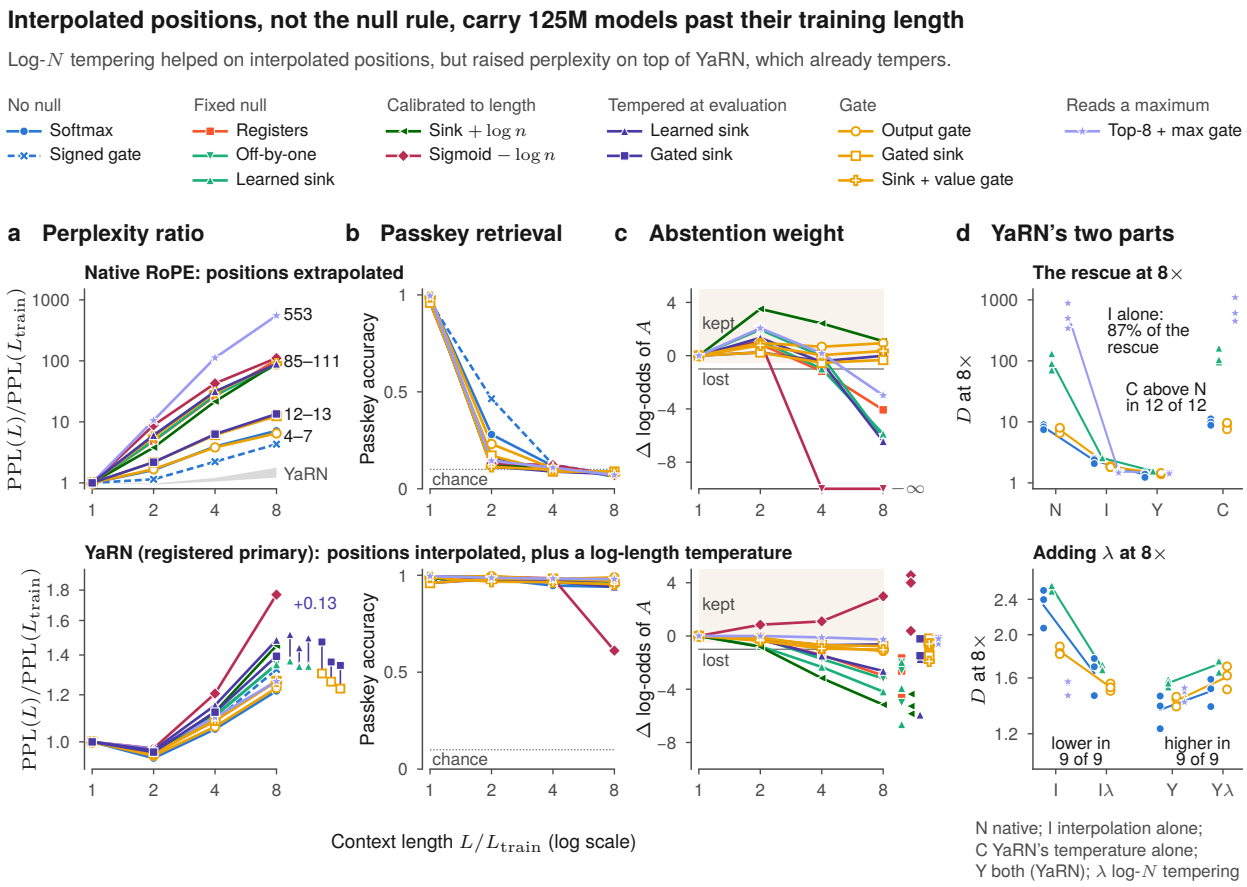


Figure 8: Interpolated positions, not the null rule, carry 125M language models past their training length, and tempering every logit by $\log N / \log N_{\text{train}}$ on top of YaRN, which already carries a log-length temperature, did not keep abstention and raised long-context perplexity. Lines join the means of 3 seeds per rule. (a) The perplexity ratio D on 95 PG-19 books; the grey band (top) spans every rule with YaRN. Tempering the same checkpoints raised D at 8× from 1.35 to 1.48 (learned sink) and from 1.27 to 1.39 (gated sink), by 0.10 to 0.17 in every seed (bars right of 8×). (b) Passkey accuracy; chance is 0.1. (c) The change from 1× in the log-odds of A, the attention mass a rule withholds from the context on the probe's absent queries (its null's mass, sigmoid's row deficit, or the share a gate closes); lines average each seed's change, as registered. Shading and line mark the registered bounds for keeping it (at least −0.5) and for losing it (at most −1), read under YaRN; marks right of 8× are seeds. At 125M no retrieval head reached the registered identification threshold (largest weight 0.247 against 0.3), so these curves are attention weights, not behaviour. Softmax and the signed gate have no null and are omitted. (d) A prespecified re-evaluation tests four rules' checkpoints at 8× on the first 24 of those books (marks are seeds). Interpolation alone (I) gives 87% of YaRN's rescue in $\log D$ (median; 70% to 99% by rule); the schedule of Theorem 8 (λ) lowers D on interpolated positions in 9 of 9 checkpoints and raises it on top of YaRN in 9 of 9; YaRN's temperature alone (C) raises it above native in 12 of 12. Models at 350M are not yet measured.

Leg	Positions	Prespecified	Median slope	95% interval	Heads
Dense	$[256, 4096)$	≤ -0.5	-0.202	$[-0.220, -0.173]$	700
Banded	$[256, 4096)$	$\lvert\cdot\rvert < 0.2$	-0.019	$[-0.030, -0.008]$	759
Banded, filling	$[8, 128)$	≤ -0.5	-0.222	$[-0.243, -0.202]$	759

Table 6: gpt-oss-20b: slopes of the document-mean sink log-odds on $\log(i+1)$, pooled over the 64 natural documents. Medians over sink-using heads, with 95% intervals from 2,000 bootstrap resamples of documents, stratified by corpus. The failed books control makes all three rows descriptive.

should stay flat, and while it fills they should fall like the dense layers. Wang (2026) reports the model's sink mass averaged over positions, 0.48 in its banded layers and 0.34 in its dense ones (their Table 14); the test here asks how the sink's log-odds change with position.

Protocol. The analysis was frozen before any production capture: the frozen bundle (SHA-256 `bd82a560...`) is dated 30 September 2026, 17:04 UTC, eight minutes before the first capture. It was not publicly deposited (Section 1). The documents are 32 PG-19 test books and 32 CPython standard-library files, each read as its first 4,096 tokens, once in natural order and once with all of those tokens shuffled. Each head's $y_h(i)$ is computed exactly, in float32, from the layer's own queries and keys (Appendix C). The sink-using heads are those whose mean sink probability exceeds 0.05 at positions 8–64 of the natural documents. For each such head, we fit the document-mean log-odds on $\log(i+1)$. The three legs apply the following prespecified criteria to the median slopes:

- dense layers at positions $[256, 4096)$, with a median slope of at most -0.5;
- banded layers at positions $[256, 4096)$, with a median slope strictly within ± 0.2;
- banded layers while the window fills, at positions $[8, 128)$, with a median slope of at most -0.5.

All three holding would confirm the law. The dense leg failing while the two banded legs hold would mean that gpt-oss compensates. Anything else is inconclusive. Before any leg is read, a control requires the natural documents' mean next-token loss to beat the shuffled documents' by more than 1 nat in each corpus, so that a broken load cannot pass.

Result. The test is undecided by its prespecified rule; descriptively it leans toward compensation. Every implementation and fidelity check passes in every cell: the forward pass is bit-identical to the model's own, and the sink probabilities from the float32 log-odds agree with the bfloat16 path's to median differences of at most 2.6×10^{-4}, against the frozen tolerance of 10^{-3}. The code corpus passes the control, with a gap of 3.87 nats, but the books corpus does not. Its gap is 0.672 nats: 5 of the 32 books have a negative gap (their loss is lower shuffled than in natural order), and 11 exceed 1 nat. The natural books text itself scored 7.06 nats per token, against 3.39 on code, far above what a model of this size reaches on ordinary prose. That leaves the prespecified outcome undecided. An audit of the text and tokens of all 64 documents found no loading defect: every document reproduces its frozen text and token-ID hashes, and every decode and shuffle check passes. It does not explain why the books control fails. The slopes would not have decided it either (Table 6). The dense leg fails, so the confirmation branch is not reached. The banded leg beyond the window holds, but the banded leg while the window fills fails, so the compensation branch is not reached either. Books and code agree (-0.20 and -0.19 on the dense leg). Selecting heads on half the documents and fitting on the other half moves no median by more than 0.012. Shuffling the tokens leaves the least-squares slopes about as flat (-0.15, 0.02 and -0.15), although the slope between query positions 256 and 4,096 is steeper for shuffled text.

What the measurement shows. Descriptively, the sink log-odds fall far more slowly than a stationary background allows, in the dense layers and in the banded layers while the band fills. In dense layers the mean background level $m(i)$ drops by about 3.4 nats between positions $[1, 10)$ and $[1000, 4096)$ of the books (3.6 in code), and the implied background mass e^{m-b}, measured per key relative to the sink, falls 30- to 35-fold. The first token takes only 3–4% of the content attention at those positions, so the flatter slope comes with a falling background, not with the first token absorbing the row. The prespecified design did not decide which mechanism lowers the background. The captured rows split the growth of the background log-sum

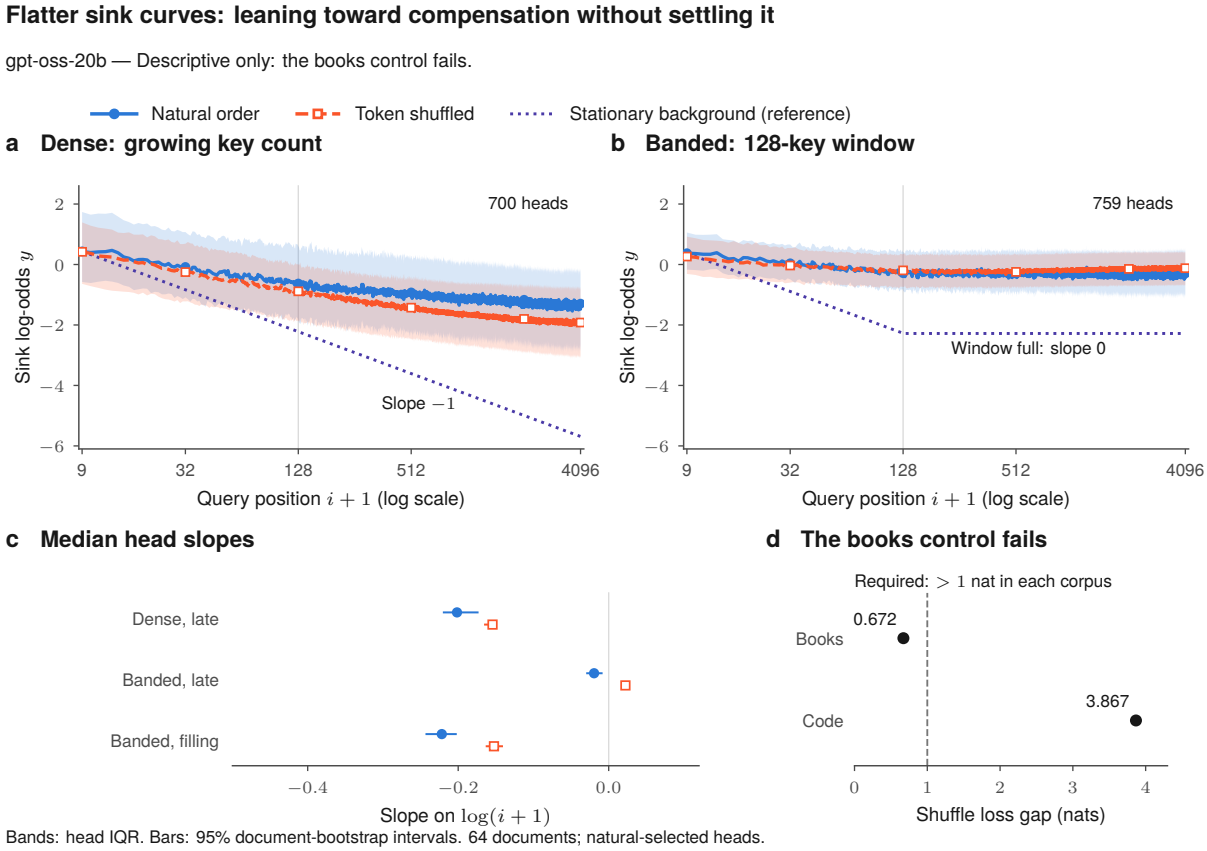


Figure 9: gpt-oss curves depart from the stationary-background reference; the prespecified test is undecided, though descriptively the slopes lean toward compensation. (a,b) Median and interquartile range of document-mean sink log-odds across sink-using heads (counts shown), selected on 64 natural documents, 32 per corpus, and reused for the shuffled condition; positions $i \in [8, 4096)$. Dotted references are $y(8) - \log[n(i)/n(8)]$, anchored to each natural median at $i = 8$, with $n(i) = i + 1$ (dense) or $\min(i + 1, 128)$ (banded). (c) Median per-head slopes, not slopes of the median curves: late positions $[256, 4096)$ and filling positions $[8, 128)$. Bars are saved 95% intervals from 2,000 corpus-stratified document-bootstrap resamples with head reselection. (d) Shuffled-minus-natural next-token loss gaps; the books gap falls below the required 1 nat. All measurements are descriptive: neither confirmation nor demonstrated compensation.

exactly, head by head. Between 256 and 4,096 keys, both contributions are measured per e-fold of the key count. Adding the far keys to the row raises its log-sum by only 0.35 to 0.56, against 1 for exchangeable keys. The log-sum of a fixed-size window of the most recent keys drifts by -0.17 to -0.31 as the query moves on. These are medians over the 700 sink-using dense heads of each corpus and condition. The captured log-sums reproduce the measured log-odds to within 6×10^{-7}. At the start of that range the window is the whole row, first token included; at its end the first token lies outside it. A distance kernel and a drift of the scores with position both hold the background down, the form that compensation through position takes (Section 5). The reading is descriptive: the prespecified test did not reach its compensated branch, and nothing here says whether the pattern would survive a passing control.

What follows. The production-model test neither confirms the fixed-sink law in gpt-oss nor, by its prespecified branches, shows that gpt-oss compensates. It is also no evidence against the identity above, whose slope of -1 requires a stationary background; gpt-oss's background is not stationary in position. This paper's claims about the fixed-sink law in trained models therefore rest on the controlled pilot (Section 6), the trained harness (Section 8) and the small trained models of Section 7.

10 Retrieval's budget on length and resolution

Retrieval is the better-known half of the law. Scaling the logits with $\log N$ keeps attention sharp on long inputs (Chiang & Cholak, 2022; Su, 2021), and the same budget appears on the resolution axis of vision models (Wang et al., 2024; Leroy et al., 2024; Gao et al., 2024); Section 2 reviews this work. We collect the program's own measurements of this half briefly. They test the retrieval side of Theorem 4 in trained models, through accuracy rather than attention weight, and they show that the retrieval rescale is one half of the tempered rule of Sections 5 and 6.

Length. A discrete retrieval task cannot test the law here: none of the 48 models trained on it under two protocols learned it, in-length retrieval was at most 2.1% against a learnability gate of 90%, and every recorded logit gap was negative. The test therefore uses a continuous one-hop retrieval task with a learned decoder. Fifteen models were trained at $N_0 = 32$, 64 and 128 keys, five seeds each, and evaluated on a doubling grid of lengths. Retrieval is the fraction of queries whose decoded output is closest in cosine to the partner's content, and its onset is the first length at which that fraction falls below 0.5. The native onsets are 8,192, 32,768 and 131,072 keys, the same in every seed at each training length. They grow as $e^{2.14\, g_0}$ in the trained logit gap g_0, which ranges from 4.18 to 5.54 (base-2 slope 3.08, descriptive 95% bootstrap interval [3.03, 3.15] over the fifteen models, resampled within three tight clusters, one per training length; Figure 10a). These are accuracy onsets of a learned decoder. The records hold no attention weights, so no weight onset is claimed, and the figure's $k = 1$ line is only a worst-case comparator. Multiplying the logits by $\log N / \log N_0$ at evaluation keeps retrieval at 0.9998 or above to 131,072 keys in all fifteen models. The native readout there has fallen to at most 5.0×10^{-5}, to 0.0007–0.0011 and to 0.33–0.46 for the three training lengths. No rescaled onset was observed, so how far the rescale holds beyond 131,072 is unknown.

The same fifteen models were first evaluated only to 16,384 keys; that run is an exact prefix of this one, not a replication. A third arm built the scale into the forward pass, but its factor is one at N_0, the only training length. Its fifteen models therefore have the same weights as these. The arm repeats the evaluation rescale's readout: 0.9999 or above to 16,384 keys (Figure 10b). It is not evidence about training with the scale.

Resolution. Vision transformers trained from scratch at one resolution, five seeds for each dataset, lose on average 43.5 and 29.6 percentage points of accuracy at the highest test resolution on CIFAR-100 and Tiny-ImageNet. Their higher test resolutions are the test images upsampled bicubically, with the position embeddings interpolated, so they add tokens but no image content. Multiplying the logits by $\log N / \log N_0$ without retraining recovers 0.65 [0.37, 0.89] and 0.83 [0.66, 1.00] points on average across test resolutions above the training resolution (95% intervals over seed–resolution pairs), 3–5% of the deficit. Building the budget into multi-resolution training gains 6.3 and 6.4 points on average at the top resolution over the same training without it (five seeds a side).

Pretrained ViT-T, ViT-S and ViT-B, evaluated once on a fixed subset of 5,000 ImageNet validation images, lose less between 224 and 384 pixels (4.56, 2.04 and 0.80 points). There the training-free rescale recovers 3.02, 1.95 and 0.12 points on average across test resolutions above 224 pixels and stays within 0.26 points of the best constant chosen for each resolution. These runs normalized their inputs with ImageNet's channel statistics, whereas the checkpoints expect a mean and standard deviation of 0.5, which depresses their accuracy, at 224 pixels included; the drops, gains and multipliers quoted for them carry that caveat.

The optimal logit multiplier c^* rises with $\log N$ (Figure 10c). Over the 130 model-resolution pairs at which models degrade gracefully, its slope in $\log N / \log N_0$ is 0.77 [0.59, 0.95] (a 95% interval over these pairs, not over models). Pooled over these pairs, $\log N$ fits better than a form linear in N (R^2 of 0.39 against 0.27). But when fitted model by model, it fits better in only 10 of the 23 models. The slope depends on the recipe: 1.05–1.13 for the pretrained ViTs, 1.40–1.42 with the budget trained in, and 0.34–0.42 for plain multi-resolution training, the last two as five-seed means on each dataset. All ten models trained from scratch at one resolution collapse at three of their six test resolutions; these models are excluded from the fit, and the registered probe of the collapse regime's boundary was inconclusive.

Two predictions that did not hold. The best fitted constant temperature matched or beat the $\log N$ rule with a fitted exponent, $(\log N / \log N_0)^p$, on extrapolated accuracy. Each was fitted at the first test resolution

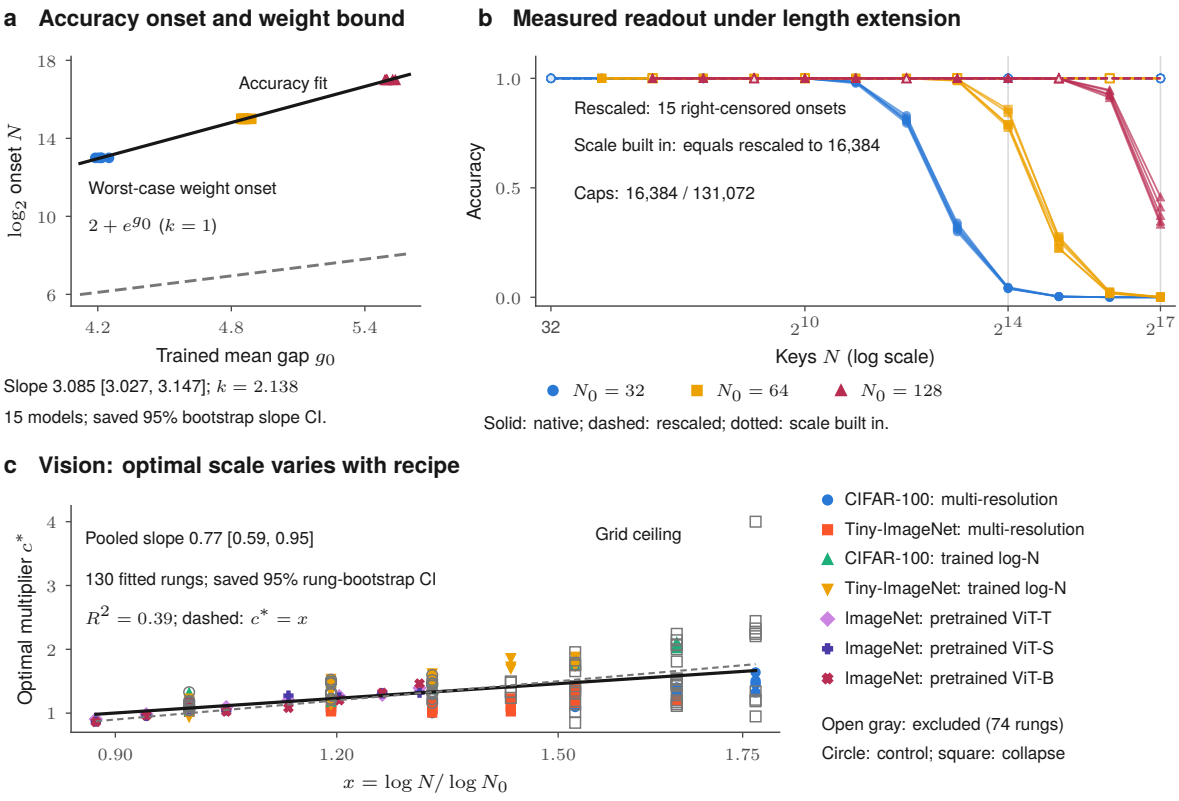


Figure 10: Retrieval accuracy fails later at larger trained gaps, and log-length rescaling preserves it to the synthetic evaluation cap; vision's optimal scale has a recipe-dependent trend. **(a)** The 15 native accuracy onsets of the continuous retrieval task (five seeds at each training length $N_0 = 32, 64, 128$), defined by the first doubling-grid value with cosine nearest-content decoding accuracy below 0.5. The trained gap g_0 is the query-mean partner logit minus the largest competing logit, excluding self. The black line uses the saved fit $\log_2 \widehat{N} = 3.0846 g_0 + 0.0028$; $k = 2.1381$ is its exponential rate. The dashed $k = 1$ comparator is the *worst-case weight onset* $2 + e^{g_0}$, not an observed weight threshold or an accuracy certificate for a mean gap. **(b)** Every seed's native (solid) and evaluation-rescaled (dashed) readout, with logits multiplied by $\log N / \log N_0$. All 15 rescaled onsets are right censored at 131,072; survival beyond that cap is unobserved. The evaluation to the lower cap is an exact prefix of the one to the higher cap, so the two marked caps describe one set of 15 models. Dotted curves show an arm with the scale built into the forward pass, tested only to 16,384; its factor is one at the training length, so it has the same 15 models' weights and repeats the rescaled readout. All 45 seed lines are retained; near-coincident curves overlap, and rescaled markers are spaced along the recorded vertices. **(c)** Saved, sub-grid-refined optimal logit multipliers: 30 from-scratch models (two datasets, three recipes, five seeds), whose higher resolutions are upsampled test images, and three pretrained ViTs, whose inputs used ImageNet normalization where their checkpoints expect a mean and standard deviation of 0.5. The pooled fit uses 130 graceful rungs from graceful sources; dashed $c^* = x$ is a reference. Open gray squares show 44 excluded collapse rungs; open circles show 30 non-collapsed rungs excluded because their source uses single-resolution training. One excluded optimum reaches the multiplier grid ceiling $c = 4$. Graceful means native accuracy at least 0.4 times in-distribution accuracy and greater than twice chance. Displayed intervals are the saved descriptive 95% bootstrap slope intervals: 5,000 model resamples in (a) and 10,000 rung resamples in (c), not a model-level interval in (c) or a pointwise confidence band. The registered probe of the low-optimum collapse regime remains inconclusive. Sources: the retrieval and multiplier records (Table 16); the from-scratch models' training records supply checkpoint provenance only.

above the training resolution and scored at the higher ones: the fitted $\log N$ rule minus the constant was -0.44 points on average over the 10 models trained at one resolution (95% interval $[-0.83, -0.04]$). As registered, that makes the budget on this axis a tunable temperature with no privileged shape. The unfitted rule, $p = 1$, matched the fitted constant, with a mean difference of 0.01 points. And the budget did not carry over to the minimal block of Bruce (2026d), an encoder built around symmetric attention scores, an independent value map, a residual update scaled by one over the depth and a register token: in models trained at one resolution, where its factor is one during training, switching it on changed top-resolution accuracy there by -0.28 and -0.21 points, against $+0.18$ and $+0.44$ for the ViT (means over five seeds), so this test measured the evaluation-time rescale. Two further tests could not be read: the heads of ViT-Tiny were too uniform for the per-head test (normalized entropy range 0.050, below its gate of 0.1), and no run reached the collapse-prone regime that the logit-magnitude test required.

One budget covers retrieval and abstention. Without a null, Theorem 4 reduces to retrieval alone: the match must outweigh the background, $e^{s_*} \geq W_* B_N$, so against background scores of a fixed level the margin must grow like $\log N$. A head that may circulate can route at that logarithmic price. A bidirectional head over N tokens that has no circulation, that is a reversible one, and that puts weight at least w_* on each token's successor needs a range of scores, the largest minus the smallest, of at least $\frac{1}{2}[(N-2)\log W_* - \log N]$, which grows linearly in N when $w_* > \frac{1}{2}$ (Appendix A; see also Bruce, 2026a, Theorem 10.1(c)). The evaluation rescale above supplies the logarithmic growth, since it multiplies every logit by $\lambda = \log N / \log N_0$. That is the tempered rule of Section 5 in a model that has no null. Once a null is present, tempering multiplies it too, and with separated margins it then keeps both halves (Theorem 8); in the pilot's isotropic geometry it kept both empirically over the tested lengths, without the certificate, and in the extension it lost abstention with heavy-tailed keys and, at the smaller budget, with clustered keys and with four 32-dimensional heads (Section 6). The results of this section are the retrieval half of the law; Sections 6 and 7 give its abstention half.

11 Recommendations and limits

Give attention a null, and set it for the length it will meet. In the 125M harness every rule that gave attention a null removed the first-token sink and cut massive activations by 67% to 82%, and, sigmoid aside, lowered perplexity at the training length by 1.2 to 2.5 percent (three-seed means against plain softmax; Section 8). A fixed null's log-odds against its background fall with slope $-\gamma$ in $\log N$, where γ is the background's growth (Section 5), so a null that must hold beyond the training length should have its logit rise like $\gamma \log N$. A stationary background with a finite exponential moment has $\gamma = 1$, and in byte models without positional encoding, adding $\ln(n/N_{\text{train}})$ to the null held its odds nearly level beyond the training length while moving the loss by less than 10^{-4} bits per byte (Section 7). The null can also live outside the scores. Natively at 8 times the training length, the output gate's perplexity ratio to its own training-length perplexity was 7.44 against plain softmax's 8.25, while a learned sink in the scores reached 97.5 (three-seed means on the first 24 books of the re-evaluation; on all 95 books, 6.47, 7.15 and 86.8; Section 8).

Keep positions in distribution when extending a model. With rotary positions, the tested causal language models compensate for a fixed null through position, and the compensation holds while positions stay in the trained range (Section 5). Most of the length rescue is positional too: interpolating the rotary frequencies alone carries a median 87% of YaRN's rescue at 8 times the training length (Section 8).

Temper where positions were interpolated; do not temper on top of YaRN. Scaling every logit, the sink's included, by $\log N / \log N_{\text{train}}$ helped where positions were interpolated: in vision transformers at higher resolution, on upsampled crops and on resized originals alike, and in causal language models on interpolated rotary positions, where it lowered the loss beyond the training length in 9 of 9 checkpoints (Section 8). On top of YaRN, which already carries a log-length temperature, it raised the loss in 9 of 9, and without interpolation YaRN's temperature alone raised long-context perplexity in 12 of 12. In these models sharpening helped only on interpolated positions, and there YaRN's own log-length temperature still beat the schedule in 9 of 9 checkpoints (Section 8), so the schedule should not replace it. The tempering guarantee needs one common multiplier on every score and separated margins (Theorem 8); none of the

Setting	Positions / comparison	Measure	Effect	Sec.
ViT-S, crops	Interpolated	Accuracy (pp)	Sink / gate with sink: 640 px: +3.5, +3.4 896 px: +4.9, +5.0 Gain in 6/6 checkpoints at each size	8
ViT-S, originals 896 px	Interpolated	Accuracy (pp)	Softmax: +4.87 [4.29, 5.46] Sink: +4.65 [3.94, 5.41] Gate + sink: +4.54 [3.93, 5.14]	8
125M LM	Interpolated $I_\lambda - I$	NLL (nats/token)	Lower by 0.20–0.50 9 of 9 checkpoints	8
125M LM	Interpolated Y vs. I_λ	NLL (nats/token)	Lower with YaRN in 9 of 9 checkpoints	8
125M LM	YaRN $Y_\lambda - Y$	NLL (nats/token)	Higher by 0.08–0.18 9 of 9 checkpoints	8
125M LM	Native $C - N$	Perplexity ratio D	Higher in 12 of 12 checkpoints	8
0.6M byte LM	Interpolated $Y - I$	Length penalty (bpb)	+0.038 to +0.075 (group means) 40/40 checkpoints worse	7
Exact-match pilot	No positional encoding	Null and match weights	Both kept in 8/8 heads 16–65,536 keys; no certificate	6

Table 7: Tempering has different effects across positional regimes. Each effect compares the same checkpoint with and without the stated temperature. Vision uses one ViT size on ImageNet-100, 3 seeds per rule, and interpolated positions. Crop gains are descriptive at 640 px and prespecified at 896 px; counts are checkpoints. Resized-original gains have recorded 95% paired hierarchical bootstrap intervals (seeds, then matched images); most originals are also upscaled. pp denotes percentage points. The 125M comparisons use 3 seeds per rule: $\lambda = \log N / \log N_{\text{train}}$ on interpolation (I) or YaRN (Y), and YaRN's temperature alone (C) on native positions (N). NLL ranges span checkpoints for tokens beyond training in 8,192-token windows; D is the full-window perplexity ratio to the training-length baseline. The $I_\lambda - I$ and $Y_\lambda - Y$ NLL comparisons are prespecified; the Y versus I_λ comparison and the $C - N$ count are descriptive. The byte-model range spans 8 position/rule means (5 seeds each; RoPE or partial RoPE): YaRN (Y: interpolation plus its temperature) versus interpolation alone (I), at 2,048 bytes. The length penalty uses a common native 256-byte chunk baseline; bpb means bits per byte. This comparison is descriptive; the temperature scales content scores, leaving learned nulls unscaled. The exact-match pilot uses 2 seeds in each of 4 configurations; retention is empirical, with no separated-margin certificate. That row covers the pilot's isotropic geometry: in the extension, tempering lost abstention with heavy-tailed keys at both budgets and, at the smaller budget, with clustered keys and with four 32-dimensional heads. Ranges are not confidence intervals; the measures have no common effect scale.

probe's retrieval heads at 125M met the margin conditions. In the pilot's isotropic geometry tempering kept both abstention and retrieval over the tested lengths, without its certificate; in the extension it lost abstention with heavy-tailed keys at both budgets and, at the smaller budget, with clustered keys and with four 32-dimensional heads (Section 6). Table 7 collects the measured effects by positional regime.

Retrieval needs a growing margin or a maximum. A null or a gate decides abstention; the match must still outweigh a background that grows with N (Theorem 4), and in the trained one-hop retrieval models of Section 10 the evaluation rescale supplies that growth. A head that reads a maximum needs no schedule against backgrounds bounded below its threshold (Section 5). In the pilot's geometries the key check, which reads the maximum for both decisions, kept abstention and retrieval except with heavy-tailed keys (Section 6). In the 125M harness, with YaRN, the head that reads a maximum was the only rule whose

null and match weights both changed by less than half a nat of log-odds out to eight times the training length, on the seed means though not in every seed (Section 8). With bounded content logits, sigmoid attention with a $-\log N$ bias holds its row mass roughly fixed but pays with single-key retrieval, whose weight falls like $1/N$ (Proposition 6); in the harness, sigmoid attention raised perplexity at the training length by 2.8 percent.

Patches within sum-to-one. Registers and value-null registers work inside unit row sums. Four registers lowered content-sink mass in a causal language model but not detectably in a ViT, and in a causal byte-level model a certified value-null register cut it by only 14%, not detectably more than a plain register token (Section 7; in the causal models the readout is attention to detector-selected tokens with small value norms).

Limits. The pilot is a mechanism-level toy without representation learning: one head and exact matches, with an extension that varies the keys, the matches, the values and the number of heads but not the model (Section 6). The main length tests are small: language models of 125M parameters with 3 seeds, byte models of about 594,320 parameters, and one ViT size on one dataset, whose higher resolutions are mostly upsampled images: resized originals were evaluated once, and most of them are themselves upscaled at 896 px. At 125M no retrieval head reached the registered identification threshold, so the abstention results there describe attention weights, not behaviour (Section 8), and the production-model test on gpt-oss was undecided by its prespecified rule, though descriptively its slopes lean toward compensation through position (Section 9). A causal softmax head has its own absorbing token (Section 4), so the decoder statements do not carry over to bidirectional attention unchanged. The retention certificate held for the cache-eviction policy based on absorbed mass. But the policy reproduced the full cache's final-layer states less accurately on average than H2O, TOVA and StreamingLLM in a linear model built from captured attention (Section 7). And how abstention depends on a head's temperature, rather than on the key count, is not treated here.

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

Habibullah Akbar. ATMA: Long-context language modeling via polar attention and gated-delta compression memory. arXiv preprint arXiv:2606.25156, 2026.

Ben Anson, Xi Wang, and Laurence Aitchison. Scale-invariant attention. In *Advances in Neural Information Processing Systems*, 2025.

Jinze Bai et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023.

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. The transformer as a constrained equilibrium: A lattice of protections for adaptive sequence mixing, and where the room is. Zenodo, released with this paper, 2026d. URL https://doi.org/10.5281/zenodo.23542051.

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

Giuseppe Bruno, Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Scaling limits of long-context transformers. arXiv preprint arXiv:2605.08505, 2026.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. In *International Conference on Learning Representations*, 2026.

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.

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

B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24 (5):2613–2626, 1981. doi: 10.1103/PhysRevB.24.2613.

James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *The Annals of Applied Probability*, 1(1):62–87, 1991.

G. David Forney, Jr. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Transactions on Information Theory*, 14(2):206–220, 1968.

Peng Gao et al. Lumina-T2X: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. arXiv preprint arXiv:2405.05945, 2024.

Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: A unified theory of signal propagation at initialisation. In *International Conference on Learning Representations*, 2026.

Siddharth Gollapudi, Nilesh Gupta, Prasann Singhal, and Sewon Min. Can language models actually retrieve in-context? Drowning in documents at million token scale. arXiv preprint arXiv:2607.01538, 2026.

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.

Tomohiro Hayase and Ryo Karakida. A unified framework for critical scaling of inverse temperature in self-attention. arXiv preprint arXiv:2605.12697, 2026.

Xingyue Huang, Xueying Ding, Mingxuan Ju, Yozen Liu, Neil Shah, and Tong Zhao. Threshold differential attention for sink-free, ultra-sparse, and non-dispersive language modeling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2026.

Anja Janßen. Limit laws for power sums and norms of i.i.d. samples. *Probability Theory and Related Fields*, 146:515–533, 2010. doi: 10.1007/s00440-008-0198-y.

Vincent Leroy, Jerome Revaud, Thomas Lucas, and Philippe Weinzaepfel. Win-win: Training high-resolution vision transformers from two windows. In *International Conference on Learning Representations*, 2024.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Physical Review Letters*, 132(7):077301, 2024. doi: 10.1103/PhysRevLett.132.077301.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Blog post, https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025.

Evan Miller. Attention is off by one. Blog post, https://www.evanmiller.org/attention-is-off-by-one.html, July 2023.

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

OpenAI. gpt-oss-120b & gpt-oss-20b model card. Model card, https://openai.com/index/gpt-oss-model-card/, August 2025. Released 5 August 2025; also arXiv:2508.10925.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state RNNs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 18724–18741, 2024.

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.

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.

Claude E. Shannon. Probability of error for optimal codes in a Gaussian channel. *Bell System Technical Journal*, 38(3):611–656, 1959.

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.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling*, 2024.

Shawn Tan, Songlin Yang, Aaron Courville, Rameswar Panda, and Yikang Shen. Scaling stick-breaking attention: An efficient implementation and in-depth study. In *International Conference on Learning Representations*, 2025.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 10347–10357, 2021.

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.

Richard Zhe Wang. Abstention and noise filtering: Two missing primitives of softmax attention. arXiv preprint arXiv:2609.22005, 2026.

ZiDong Wang, Zeyu Lu, Di Huang, Cai Zhou, Wanli Ouyang, and Lei Bai. FiTv2: Scalable and improved flexible vision transformer for diffusion model. arXiv preprint arXiv:2410.13925, 2024.

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.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, 2023.

Zayd Muhammad Kawakibi Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. Softpick: No attention sink, no massive activations with rectified softmax. arXiv preprint arXiv:2504.20966, 2025.

A Proofs

Every finite identity and bound below was also checked numerically on random float64 instances; the checks are falsification tests, not proofs. Logarithms are natural.

A.1 The constraint (Section 3)

Proof of Proposition 1. Differentiating $p_j = e^{s_j}/(e^b + \sum_k e^{s_k})$ gives $\partial p_j/\partial s_k = p_j(\delta_{jk} - p_k)$, that is $J = \operatorname{diag}(p) - pp^\top$, which is also the Hessian of $\log(e^b + \sum_k e^{s_k})$. For $x \in \mathbb{R}^N$, $x^\top Jx = \sum_j p_j x_j^2 - (\sum_j p_j x_j)^2$, and Cauchy–Schwarz with weights p_j gives $(\sum_j p_j x_j)^2 \le r \sum_j p_j x_j^2$, hence $x^\top Jx \ge (1-r)\sum_j p_j x_j^2 \ge (1-r)\min_j p_j \|x\|^2$. Finally $p = r\,\mathrm{softmax}(s)$ with $r = \sum_j e^{s_j}/(e^b + \sum_j e^{s_j}) = \sigma(\mathrm{LSE}(s) - b)$. □

For a row-stochastic P with stationary law π, Jensen's inequality gives $(Pv)_i^2 \le \sum_j P_{ij} v_j^2$, and summing against π with $\pi^\top P = \pi^\top$ gives $\|Pv\|_\pi^2 \le \|v\|_\pi^2$, with equality on constants.

A.2 Gate times routing and the Markov reading (Section 4)

Proof of Proposition 2. $P_{ij} = \dfrac{\sum_k e^{s_{ik}}}{z_i + \sum_k e^{s_{ik}}} \cdot \dfrac{e^{s_{ij}}}{\sum_k e^{s_{ik}}}$, and the first factor is $\sigma(\mathrm{LSE}_i(s) - \log z_i)$. □

Proof of Theorem 3. Perron–Frobenius supplies positive eigenvectors, and a positive root because Q is irreducible and nonzero. Multiplying $Q\mathbf{1} \le \mathbf{1}$, strict in some row, by the positive left eigenvector gives $\rho\nu^\top\mathbf{1} < \nu^\top\mathbf{1}$, hence $\rho < 1$. The equation $Qh = \rho h$ gives unit row sums for P^h. Summing $\nu_i h_i P^h_{ij} = \nu_i Q_{ij} h_j/\rho$ over i proves stationarity of $\pi^h \propto \nu \circ h$. The ratio $\pi^h_j P^h_{ji}/\pi^h_i$ equals $\nu_j Q_{ji}/(\rho\nu_i)$, which gives the reversal. Jensen's inequality and stationarity give $\|P^h\|_{L^2(\pi^h)} = 1$, and the diagonal similarity gives $\|Q\|_w = \rho$. For $Q_{ij} = e^{s_{ij} - L_i}$ the forward row normalizer is $\sum_j e^{s_{ij}} h_j = e^{L_i}\rho h_i$ and the reverse normalizer is $\sum_j e^{s_{ji}}\nu_j e^{-L_j} = \rho\nu_i$, which give the two softmax forms. With structural zeros the same formulas hold on the support. □

Strict killing in some row is not enough without irreducibility: a disconnected stochastic block can keep eigenvalue one. Uniform survival at most $r < 1$ in every row gives $\|Q^t\|_\infty \le r^t$ directly.

Proposition 12 (Endpoint and occupation limits)**.** *If Q is also primitive and $\nu^\top h = 1$, then for any initial law μ, conditioned on survival to time t, the endpoint law converges to $\nu/\nu^\top\mathbf{1}$, and the empirical path occupation converges in probability to $\pi^h = \nu \circ h$.*

Proof. The Perron expansion $Q^t = \rho^t h\nu^\top + o(\rho^t)$, divided by $\mu^\top Q^t\mathbf{1}$, gives the endpoint limit. For a path $x_0, \dots, x_t$, telescoping the ratios of h gives $\prod_{j<t} Q_{x_j x_{j+1}} = \rho^t\big(h(x_0)/h(x_t)\big)\prod_{j<t} P^h_{x_j x_{j+1}}$, so the survival-conditioned path law has a density proportional to $h(X_0)/h(X_t)$ relative to the P^h path with the same initial law. This density is bounded above independently of t. The law of large numbers for the primitive chain P^h concentrates its occupation on π^h, and the bounded density transfers the concentration to the conditioned law. □

Primitivity rules out periodic oscillation of the endpoint. The causal, reducible case needs its own accessibility analysis: $Q = \operatorname{diag}(0.9, 0.5)$ started in state 2 stays there whenever it survives, so conditioning cannot move it to the state with the larger diagonal, and Jordan chains can add polynomial factors. For an absorbing chain with transient block T from which every state reaches the target, take the longest of the finitely many shortest absorbing paths and the smallest of their positive probabilities: some finite m has $\|T^m\|_\infty < 1$, hence $T^t \to 0$ and all mass is absorbed.

A.3 The law (Section 5)

Theorem 4 with present keys counted. If N counts all keys of the present query rather than the distractors, replace N by $N - 1$ in the retrieval statements. The fixed-sink bounds of Section 5 follow from $B_N \ge Ne^{-\tau}$ and $e^{s_*} \le e^\tau$: $a_N = e^b/(e^b + B_N) \le \sigma(b + \tau - \log N)$ and $w_N \le e^\tau/(e^\tau + Ne^{-\tau})$.

Proof of Proposition 5. The targets read $z_0 \geq A_* B_0$ and $z_1 \leq M/W_* - B_1$, and the sensitivity bound gives $z_1 \geq e^{-\epsilon} z_0 \geq e^{-\epsilon} A_* B_0$, which proves necessity. Conversely, $z_0 = A_* B_0$ and $z_1 = e^{-\epsilon} A_* B_0$ are positive with log-ratio exactly ϵ; they meet abstention with equality and meet retrieval exactly when the condition holds. With bounded scores, $B_0, B_1 \geq N e^{-\tau}$ and $M \leq e^{\tau}$ turn the condition into $N \leq N_\epsilon$. Since $1/(1 + A_*) = 1 - a_*$, $N_0 = N_*$, and N_ϵ increases to $e^{2\tau}(1 - w_*)/w_* = N_*/(1 - a_*)$. Distractors at $-\tau$, the match at τ and these nulls meet both targets for every $N \leq N_\epsilon$, with retrieval exactly w_* at $N = N_\epsilon$. For equal backgrounds $B_0 = B_1 = B$ with $r = M/(W_* B)$, some finite ϵ works exactly when $r > 1$, and the least one is $\max\{0, \log(A_*/(r-1))\}$. For the last statement, absent success is $Z \geq A_* B$ and present success is $Z \leq M/W_* - B < A_* B$, so the event $\{Z \geq A_* B\}$ has probability at least $1 - \alpha$ under absence and at most β under presence. □

Proof of Proposition 6. The absent constraint gives $N\sigma(-\tau + c_N) \leq M$. For $N > M$ this gives $e^{c_N} \leq e^{\tau} M/(N - M)$, hence $p_* \leq e^{2\tau} M/(N - M + e^{2\tau} M)$, and $p_* \geq w_*$ rearranges to the bound. Sigmoid weights without a row-mass guarantee have no probabilistic abstention mass, and the proposition must not be applied to them under that name. □

Proof of Proposition 7. For a variable in $[l, u]$ the exponential-moment bound $\mathbb{E}e^{t(X - \mathbb{E}X)} \leq e^{t^2 (u-l)^2/8}$ holds, since the second derivative of the log moment-generating function is a variance under an exponential tilt, at most $(u - l)^2/4$, and its value and first derivative vanish at zero. Markov's inequality on the product of the moments, optimized over t and union-bounded over the two tails, gives the concentration of B_N/N, and monotonicity of $z/(z + B)$ gives the interval. Jensen gives $\mathbb{E}a_N \geq a_{\mathrm{mf}}$. Taylor's theorem with $f''(B) = 2z/(z + B)^3$, $B \geq Nl$ and $\mathrm{Var}(B_N) = Nv$ gives the upper bound. With only a finite variance, Chebyshev with $t = \sqrt{v/(N\delta)}$ replaces the exponential bound. □

Proof of Theorem 8. After scaling, the background sum is at most $N e^{\lambda s_0} = N_0^{\lambda} e^{\lambda s_0}$. Dividing by the null's $e^{\lambda b}$ gives $a_N \geq (1 + v_a^{\lambda})^{-1}$, and dividing by the match's $e^{\lambda s_*}$ gives the retrieval bound. Powers of numbers in $[0, 1]$ are non-increasing in λ, and tend to zero for numbers below one. These are sufficient conditions, not necessary ones at each finite context. □

Proof of Theorem 9. Divide the null's denominator by $e^{\lambda b}$ and use $\sum_j e^{\lambda s_j} = N M_{\lambda,N}^{\lambda}$ with $N = N_0^{\lambda}$. For $\lambda \geq 1$, Jensen applied to $x \mapsto x^{\lambda}$ gives $M_{\lambda,N} \geq M_{1,N}$, and each term is at most $e^{\lambda s_{\max}}$; the reciprocal is decreasing in its background argument. □

The identity is exact for the empirical scores. For a fixed law replicated to N keys the empirical power mean is that law's; for random scores the population power mean $M_q = (\mathbb{E}e^{qS})^{1/q}$ gives a population-sum prediction, not an exact finite-sample probability. The realized a_N need not be monotone in N, and the $b = s_0$ counterexample assumes a background exactly at s_0.

Proof of Theorem 10. Write $\lambda_N = \log N / \log N_0$ and $B_N = \sum_{j \leq N} e^{\lambda_N (S_j - b)}$. Almost surely every score is at most s_0, so $B_N \leq e^{\lambda_N (h-b)}$. For each $\delta > 0$, $p_\delta = \mathbb{P}(S > s_0 - \delta) > 0$, and the strong law for its indicators puts at least $N p_\delta/2$ scores above $s_0 - \delta$ eventually, so $B_N \geq (p_\delta/2) e^{\lambda_N (h - b - \delta)}$. Taking logarithms, dividing by λ_N and letting $\delta \downarrow 0$ along a countable sequence gives $\log B_N / \lambda_N \to h - b$ almost surely, which proves both strict cases. At $b = h$, $B_N = N^{-1} \sum_j e^{-\lambda_N (s_0 - S_j)}$: the proportion of scores equal to s_0 tends to p_0 and bounds it below, and for any $\delta > 0$ it is bounded above by the proportion above $s_0 - \delta$ plus $e^{-\lambda_N \delta}$, whose limit superior is at most $p_\delta \downarrow p_0$. Hence $B_N \to p_0$. The retrieval background term is the same normalized sum times $e^{\lambda_N (h - s_*)}$, which with $e^{\lambda_N (b - s_*)}$ gives every retrieval limit. For unbounded support use a threshold $T > b - \log N_0$ with $\mathbb{P}(S > T) > 0$ in the tail-count bound. □

Independence is used only for the convergence of the fixed-threshold frequencies, so the conclusions hold for any sequence with those limits and the same upper bound, such as a stationary ergodic one. Repeating one random score forever is a stationary, nonergodic counterexample, whose threshold is the realized score plus $\log N_0$.

Proposition 13 (Finite-sample comparison)**.** *In the bounded i.i.d. setting let* $U_{j,N} = e^{-\lambda_N(s_0 - S_j)} \in [0,1]$, $\mu_N = \mathbb{E}U_{j,N}$ *and* $v = N_0 e^{s_0 - b}$. *For* $0 < \delta < 1$ *and* $t_N = \sqrt{\log(2/\delta)/(2N)}$, *with probability at least* $1 - \delta$,

$$\frac{1}{1 + v^{\lambda_N}(\mu_N + t_N)} \le a_N \le \frac{1}{1 + v^{\lambda_N} \max(0, \mu_N - t_N)},$$

and $\mathbb{E}a_N \ge \left[1 + v^{\lambda_N}\mu_N\right]^{-1}$.

Proof. The bounded-variable bound gives $\mathbb{P}\big(|N^{-1}\sum_j U_{j,N} - \mu_N| > t\big) \le 2e^{-2Nt^2}$; use $B_N = v^{\lambda_N} N^{-1}\sum_j U_{j,N}$ and monotonicity. Convexity of $x \mapsto (1 + v^{\lambda_N}x)^{-1}$ gives the expectation bound. □

Proof of Theorem 11. For $Y = s_0 - S \ge 0$, integration by parts gives $\mathbb{E}e^{-\lambda Y} = \lambda \int_0^\infty e^{-\lambda y}\mathbb{P}(Y \le y)\,dy$. Substitute $u = \lambda y$ and multiply by λ^a. On $y \le \eta$ the tail assumption bounds the integrand by a constant times $e^{-u}u^a$, and on $y > \eta$ the integral is at most $\lambda^a e^{-\lambda\eta}$, so dominated convergence gives $\mu_N \sim C\Gamma(a+1)\lambda_N^{-a}$. The concentration argument with $t_N = \sqrt{2\log N/N}$ has summable failure probabilities $2N^{-4}$, so by Borel–Cantelli the deviation is eventually at most t_N almost surely. Since $t_N = o(\lambda_N^{-a})$, the empirical mean is asymptotic to μ_N, which tends to zero because $a > 0$, and substitution gives the three assertions. □

The successor bound of Section 10. Let P be the row softmax of scores S whose range, the largest minus the smallest, is R, and let P be reversible with stationary law π, so that $\pi_i P_{ij} = \pi_j P_{ji}$. If $P_{i,i+1} \ge w_*$ for every $i \le N-1$, then $P_{i+1,i} \le 1 - w_*$ for $i \le N-2$, so $\pi_{i+1}/\pi_i = P_{i,i+1}/P_{i+1,i} \ge W_*$ and $\pi_{N-1}/\pi_1 \ge W_*^{N-2}$. Detailed balance also gives $\pi_{N-1}/\pi_1 = (Z_{N-1}/Z_1)\,e^{S_{1,N-1} - S_{N-1,1}}$ with $Z_i = \sum_j e^{S_{ij}}$, and each factor is at most e^R. Hence $R \ge \frac{1}{2}(N-2)\log W_*$, which implies the bound stated there.

B Pilot protocol, full tables and robustness

B.1 Task, training and seeds

Task and rules. Queries and keys are isotropic unit vectors in $\mathbb{R}^{128}$. Only their cosines matter, so each background cosine is drawn exactly from the distribution of the inner product of a fixed unit vector with a uniformly random one. A present query has one key at cosine one, placed first (position plays no role), and $N-1$ background keys, one fewer than the matched convention of Theorem 4; an absent query has N background keys. Each key carries a uniformly random one-hot value among 16 classes, and the register's key direction is drawn afresh for each query. The seven rules are those of Section 6. The learned parameters are θ, through $\tau = \tau_{\max}\sigma(\theta)$, the offset β_0 and the null logit b, which softmax does not use.

Training and evaluation. Each head is trained for 800 Adam steps at learning rate 0.05, on batches of 512 queries at N_{train} keys, half present and half absent, by the squared error between o and its target. The evaluation draws 1,024 fresh queries per length and seed, about half absent, at the 13 lengths $16, 32, \ldots, 65{,}536$. It runs in float64, in chunks at long lengths. The law is evaluated exactly for isotropic keys: the logarithm of the cosine's moment generating function is checked against quadrature to 10^{-10}, and nothing is fitted. The re-analysis behind these tables first reproduces the pilot's printed summary tables from the original records, and only then computes anything new.

Seeds. The two seeds agree closely for most rules, but not for all. Of the 224 printed seed-mean entries, 13 have seeds that differ by more than 0.01, all of them for the register, plain sigmoid or softmax. The largest ranges are 0.151, in the register's learned τ at $\tau_{\max} = 20$, $N_{\text{train}} = 64$; 0.774, in plain sigmoid's abstention at 65,536 keys, a value near -670; and 0.010 for softmax. For the other four rules no entry differs by more than 0.0081. Onsets are therefore given per seed (Tables 1 and 8). Figure 11 plots the weight onsets of the six untempered rules against each head's learned scale τ, beside the limits that Theorem 4 sets on keeping both weights above one half at once; the limits do not bound either weight alone, and they do not apply to the sigmoid rules.

	Softmax	Value-null register	Fixed sink	Sink $+\log N$	Sigmoid	Sigmoid $-\log N$	$\log N$-tempered
$\tau_{\max} = 10$, $N_{\text{train}} = 64$							
$a < 1/2$, measured	$\leq$16	512	512	—	64	—	—
$a < 1/2$, law	$\leq$16	360/380	440	—	48	—	2.3×10^{12}
$w < 1/2$, measured	2,048	16,384	16,384	2,048	—	2,048	—
$w < 1/2$, law	1,600	12,000	12,000	1,500	—	1,300	—
$\tau_{\max} = 10$, $N_{\text{train}} = 512$							
$a < 1/2$, measured	$\leq$16	1,024	1,024	—	64	—	—
$a < 1/2$, law	$\leq$16	610	910	—	51	—	$8.4\times10^{6}/8.6\times10^{6}$
$w < 1/2$, measured	16,384	16,384	16,384	8,192	$\leq$16	4,096	—
$w < 1/2$, law	10,000	12,000	12,000	4,700	$\leq$16	4,000	—
$\tau_{\max} = 20$, $N_{\text{train}} = 64$							
$a < 1/2$, measured	$\leq$16	8,192/16,384	8,192	—	64	—	—
$a < 1/2$, law	$\leq$16	8,300/8,600	4,600/4,500	—	49	—	$9.4\times10^{12}/9.3\times10^{12}$
$w < 1/2$, measured	2,048	—	—	32,768	—	8,192	—
$w < 1/2$, law	1,600	$6.1\times10^{6}/5.3\times10^{6}$	1.6×10^{6}	25,000	—	7,300/7,400	—
$\tau_{\max} = 20$, $N_{\text{train}} = 512$							
$a < 1/2$, measured	$\leq$16	—	65,536	—	64	—	—
$a < 1/2$, law	$\leq$16	75,000/78,000	33,000	—	52	—	1.5×10^{14}
$w < 1/2$, measured	32,768	—	—	—	$\leq$16	65,536	—
$w < 1/2$, law	25,000/26,000	7.1×10^{7}	1.6×10^{7}	2.2×10^{5}	$\leq$16	57,000	—

Table 8: Pilot weight onsets: the first length at which the mean abstention a (absent queries) or the mean match weight w (present queries) falls below $1/2$, measured and as the mean-field law predicts from each head's learned parameters. The law's onset is continuous, to two significant figures; the measured one is a point of the grid $16, 32, \ldots, 65{,}536$. Entries show "seed 0/seed 1" where the seeds differ. — means no onset (measured: up to 65,536 keys; law: no root through 10^{30}), and $\leq$16 means below $1/2$ already at the shortest length. The tempered rule's law is the stationary power-mean law, checked numerically.

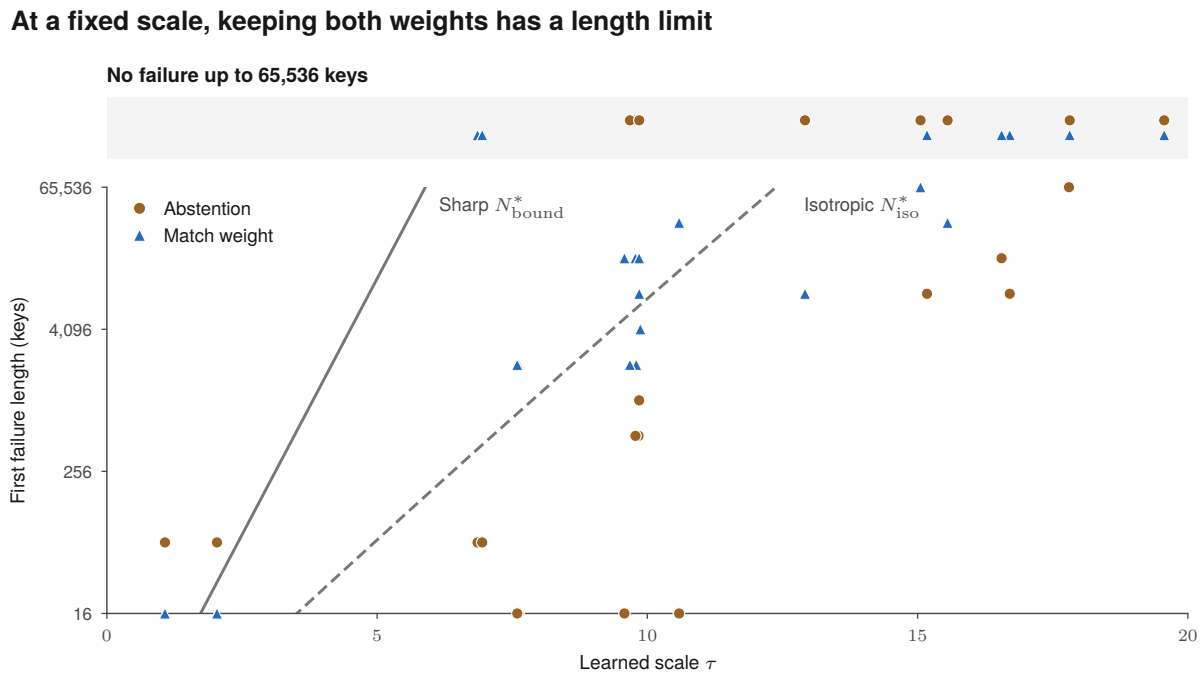


Figure 11: At a fixed logit scale, keeping both weights above one half has a length limit. The 96 marks show the six untempered rules at logit budgets 10 and 20, training lengths 64 and 512, and two seeds: brown circles are absent-query abstention, blue triangles present-query match weight. Each of the 64 marks in the main plot is the first tested length below one half, against the learned scale τ; 12 are already below one half at 16 keys. The 32 marks in the separate top band have no failure through 65,536 keys. The neutral curves are simultaneous half-weight limits for normalized exponentials with a shared null and N background keys in both queries (Theorem 4): the sharp bounded-score limit $N^*_{\text{bound}} = e^{2\tau}/2$ (solid), and the isotropic mean-field counterpart $N^*_{\text{iso}} = e^{\tau}/(2\mu(\tau))$ (dashed), where $\mu(\tau) = \mathbb{E}e^{\tau C}$ for the cosine C of isotropic keys in $\mathbb{R}^{128}$. They limit keeping both weights simultaneously, not either individual weight, and are not sigmoid bounds. The pilot's present queries have $N-1$ background keys, one fewer than the matched construction. Tempering changes the scale with N and is excluded (Figure 3). The measurements use 1,024 evaluation queries per length and seed; these are mean attention weights, not task accuracies.

B.2 The robustness extension: design and checks

Design. The extension was registered and frozen before it ran, and its reproduction stage had to pass before anything else was read. Its 927 cells are 57 reproduction cells (the pilot's seven rules in its four configurations, with its two seeds, and one cell of fixtures), 720 geometry cells and 150 mixed-length cells. The geometry cells train nine rules, the pilot's seven and two that read the largest score, at both budgets and $N_{\text{train}} = 64$, with seeds 0 to 4, in eight geometries. Each changes one factor of the pilot, except the last. *Clustered keys*: 16 clusters with sizes proportional to $1/c$ and intra-cluster cosine 0.3, drawn afresh for each query, which belongs to one of them. *Heavy-tailed keys*: each background key adds a Student-t component, with 3 degrees of freedom and scale 1, along the query. *Near matches*: the match has cosine $c_m \in \{0.95, 0.80, 0.60\}$, and every row, present or absent, holds four hard distractors at $c_m - \delta$ with $\delta \sim U[0.1, 0.3]$. *Continuous values*: values with the mean and covariance of the one-hot law, read by a decoder that must pick the target among 16 candidates. *Four heads*: continuous values read by four heads on disjoint 32-dimensional subspaces, a quarter of the pilot's head dimension, with an identity output map; a row abstains when the heads' mean mass is below 1/2. The mixed-length cells train, with isotropic, clustered and heavy-tailed keys, a fixed sink, the calibrated sink $+\log N$, a sink whose null logit adds $\gamma \log(N/64)$ with γ learned, and tempering applied in training, on lengths drawn uniformly from $\{64, 128, 256, 512, 1{,}024\}$, and a fixed sink trained at 1,024 keys.

Rule	τ	Abstention a			Retrieval acc.	Abstention acc.	Seed
		N_{train}	4,096	65,536	4,096 / 65,536	4,096 / 65,536	range
$\tau_{\max} = 10$, $N_{\text{train}} = 64$							
Softmax	7.62	0.000	0.000	0.000	1.00 / 1.00	0.00 / 0.00	0.006
Value-null register	9.84	0.831	0.109	0.009	1.00 / 1.00	0.00 / 0.00	0.011
Fixed sink	9.78	0.874	0.097	0.007	1.00 / 1.00	0.00 / 0.00	0.001
Sink $+ \log N$	9.68	0.868	0.867	0.866	0.00 / 0.00	1.00 / 1.00	0.001
Sigmoid	6.86	0.330	−42	−688	0.99 / 0.29	0.00 / 0.00	0.638
Sigmoid $- \log N$	9.79	0.895	0.893	0.893	0.00 / 0.00	1.00 / 1.00	0.001
$\log N$-tempered	9.78	0.874	0.958	0.971	1.00 / 1.00	1.00 / 1.00	0.001
$\tau_{\max} = 10$, $N_{\text{train}} = 512$							
Softmax	9.58	0.000	0.000	0.000	1.00 / 1.00	0.00 / 0.00	0.000
Value-null register	9.84	0.535	0.161	0.014	1.00 / 1.00	0.02 / 0.00	0.009
Fixed sink	9.85	0.640	0.182	0.014	1.00 / 1.00	0.00 / 0.00	0.000
Sink $+ \log N$	9.85	0.640	0.640	0.640	1.00 / 0.00	1.00 / 1.00	0.000
Sigmoid	1.07	−4.065	−40	−647	0.09 / 0.07	0.00 / 0.00	0.024
Sigmoid $- \log N$	9.87	0.701	0.701	0.701	1.00 / 0.00	1.00 / 1.00	0.001
$\log N$-tempered	9.85	0.640	0.645	0.623	1.00 / 1.00	1.00 / 1.00	0.001
$\tau_{\max} = 20$, $N_{\text{train}} = 64$							
Softmax	7.60	0.000	0.000	0.000	1.00 / 1.00	0.00 / 0.00	0.008
Value-null register	16.63	0.983	0.635	0.194	1.00 / 1.00	0.71 / 0.09	0.151
Fixed sink	15.18	0.986	0.527	0.065	1.00 / 1.00	1.00 / 0.00	0.003
Sink $+ \log N$	15.55	0.989	0.989	0.989	1.00 / 0.00	1.00 / 1.00	0.004
Sigmoid	6.94	0.347	−41	−670	0.99 / 0.30	0.00 / 0.00	0.774
Sigmoid $- \log N$	12.91	0.966	0.965	0.965	1.00 / 0.00	1.00 / 1.00	0.008
$\log N$-tempered	15.18	0.986	0.999	1.000	1.00 / 1.00	1.00 / 1.00	0.003
$\tau_{\max} = 20$, $N_{\text{train}} = 512$							
Softmax	10.58	0.000	0.000	0.000	1.00 / 1.00	0.00 / 0.00	0.010
Value-null register	19.56	0.976	0.887	0.541	1.00 / 1.00	0.95 / 0.56	0.027
Fixed sink	17.80	0.985	0.891	0.338	1.00 / 1.00	1.00 / 0.00	0.002
Sink $+ \log N$	17.81	0.986	0.986	0.986	1.00 / 1.00	1.00 / 1.00	0.000
Sigmoid	2.03	−3.913	−38	−628	0.13 / 0.08	0.00 / 0.00	0.092
Sigmoid $- \log N$	15.05	0.960	0.960	0.960	1.00 / 1.00	1.00 / 1.00	0.004
$\log N$-tempered	17.80	0.985	0.994	0.997	1.00 / 1.00	1.00 / 1.00	0.002

Table 9: The four pilot configurations in full: seed means of the learned scale, of the abstention a at the training length, 4,096 and 65,536 keys, and of retrieval and abstention accuracy at 4,096 and 65,536 keys, with each rule's largest range between the two seeds over these entries. Plain sigmoid's abstention is negative because its row mass exceeds one.

The max gate routes as softmax does and abstains through a gate $\sigma(a_g(M - \theta_g))$ on the row's largest score M; the key check uses the same gate but routes over the eight best keys, so neither of its decisions sums over the N keys.

Reproduction and checks. The fixture cell first failed two registered checks for reasons outside the code under test: one compared floating-point outputs with a digest computed on another machine, and one tested 152 moments of the continuous-value law each at three standard errors, a rule that a correct generator fails with probability 0.34. An amendment to the registration replaced them with a comparison against the pinned reference in the same process and with exact checks of the generator plus a simultaneous threshold of 4.5 standard errors (family-wise rate 10^{-3}), both fixed before the fixture and reproduction cells ran again. In all 56 reproduction cells the port of the pilot's script equals a pinned copy of the script, run in the same process, bit for bit. Against the stored pilot records, eight softmax cells differ in one parameter, the offset β_0, by at most 8.9×10^{-10}: β_0 cancels in softmax's output, so it stays at rounding level (at most 1.1×10^{-9} in size) and fails the registered relative tolerance of 10^{-5}. As registered for this case, the gate then takes the port's

Keys	$\tau_{\max}$	Onsets	Rules that sum over the keys	Tempering	Key check, max gate
Isotropic	10	85/85	—	background, 10/10	5/5, 5/5
	20	51/51	—	background, 10/10	5/5, 5/5
Clustered	10	80/80	Tempered: abstention accuracy below 0.9, 9/10; calibrated, mixed lengths: abstention drifts, 2/5	background, 10/10	5/5, 5/5
	20	50/50	—	background, 10/10	5/5, 5/5
Heavy-tailed	10	20/20	Calibrated: abstention drifts, 15/15	loses abstention, 10/10	0/5, 0/5
	20	20/20	Calibrated: abstention drifts, 15/15	loses abstention, 10/10	0/5, 0/5
Near match 0.95	10	75/75	—	background, 5/5	5/5, 5/5
	20	25/29	Register: abstention indeterminate, 1/5	background, 5/5	5/5, 5/5
Near match 0.80	10	71/75	—	background, 5/5	5/5, 5/5
	20	20/25	Register: abstention indeterminate, 2/5	background, 5/5	5/5, 5/5
Near match 0.60	10	75/75	—	background, 5/5	5/5, 0/5
	20	51/54	—	background, 5/5	5/5, 5/5
Continuous values	10	50/50	—	background, 5/5	5/5, 5/5
	20	40/40	—	background, 5/5	5/5, 5/5
Four heads	10	220/220	—	background, 20/20	20/20, 20/20
	20	180/180	—	background, 14/20; empirical, 6/20	20/20, 20/20

Table 10: The robustness extension by geometry and budget, five seeds each. Onsets: scored weight-onset comparisons whose measured onset agrees with the conditional mean-field law under the one-grid-step rule below. Rules that sum over the keys: the registered pattern checks that fail, with the cells that fail out of those checked (calibrated rules: sink $+\log N$ and sigmoid $-\log N$ trained at 64 keys and sink $+\log N$ trained on mixed lengths; tempered rules: tempering at evaluation and tempering in training). "Drifts": the mean abstention falls by more than 0.02 over the grid. Tempering: the label of each tempered row (background: explained by background concentration); none is certified. Last column: rows in which the key check and the max gate keep both sides (per head in the four-head geometry).

own records as the baseline; the eight cells are kept, marked mixed, with their differences. The assessment is mixed for that reason and one other: the advisory joint-budget check fails in 4,454 of its 31,600 per-query comparisons, 4,264 of them for the two heads that read the largest score, which lie outside the budget's class, and 190 for the calibrated and learned-length sinks trained on mixed lengths with clustered keys. Every hard check passes: softmax retrieves at the training length with accuracy at least 0.97 in every geometry, budget and seed; the tempered and fixed-sink heads of each setting have identical trained parameters; and the rules of each setting see the same evaluation stream.

B.3 The robustness extension: results

Table 10 gives each geometry and budget and Table 11 the mixed-length cells; Figure 14 shows every geometry cell at $\tau_{\max} = 10$ in sixteen plates. The registered pattern checks fail in 44 of their 920 rows: with heavy-tailed keys, the drift of every calibrated rule's mean abstention (30 rows); with clustered keys at $\tau_{\max} = 10$, the tempered rules' abstention accuracy (nine of ten rows) and the drift of the mixed-length calibrated sink in two seeds, by 0.021 against the tolerance of 0.02; and with near matches at cosine 0.80 and 0.95 at $\tau_{\max} = 20$, the register, whose abstention is neither kept within the registered band nor lost past its threshold in three cells. None of these is a refutation: no rule without a length input keeps both sides in a cell where the law puts a failure inside the grid, in any geometry. With near matches at cosine 0.95 and $\tau_{\max} = 20$, the fixed sink keeps both sides in five seeds and the register in three, where the law puts their failures beyond the grid.

Weight onsets. A comparison is scored when the law is in its concentration regime at the predicted onset (the relative standard deviation of the background sum at most 0.5, with a tail-sensitivity bracket where a few queries or rare keys dominate the moment) and at least one of the two onsets falls on the grid. It passes when the measured onset lies within one grid step of the first grid point at or above the predicted one.

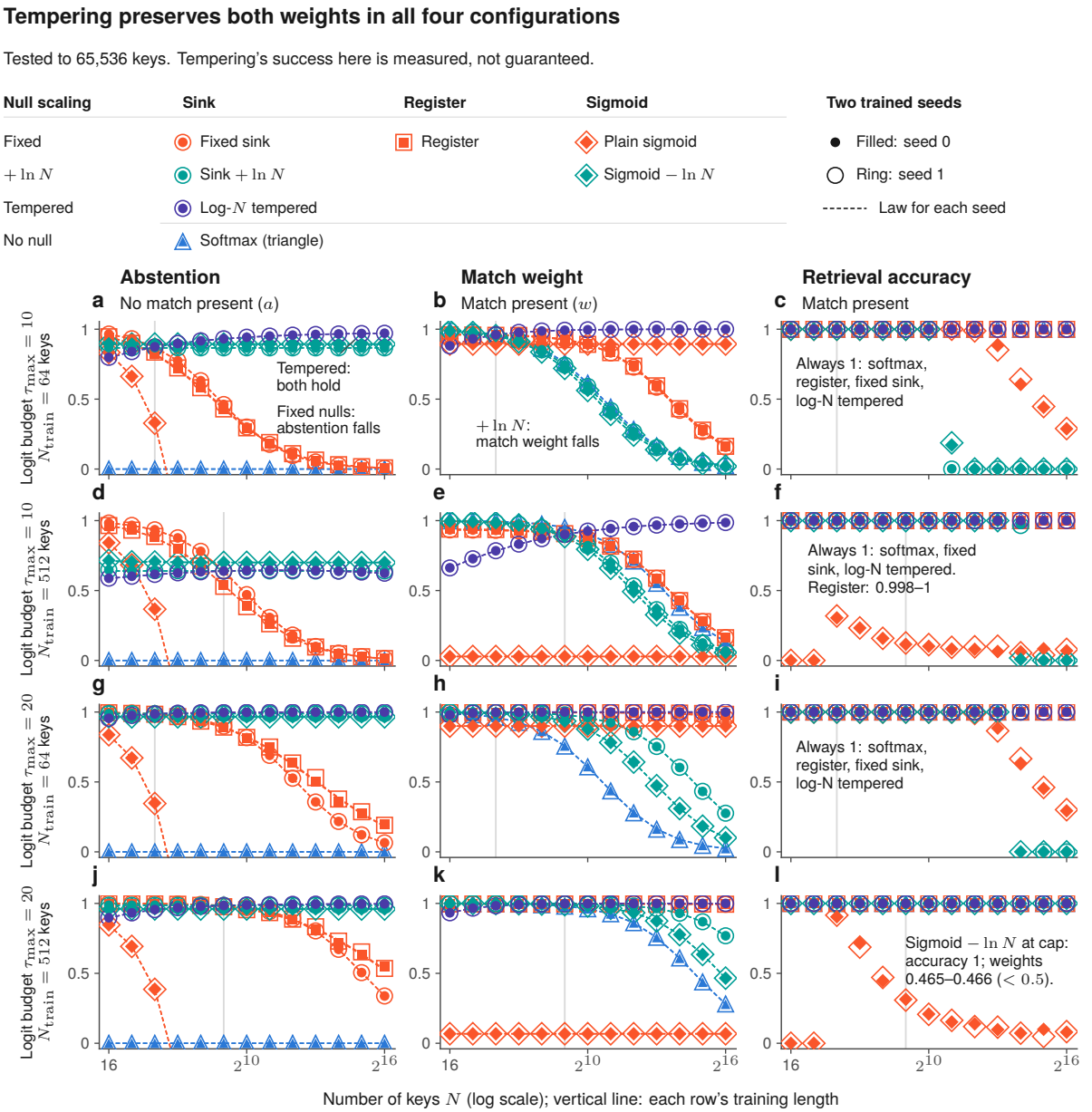
Keys	$\tau_{\max}$	Fixed sink: abstention fails from at 64	mixed	at 1,024	Learned length $\hat{\gamma}$ (keeps both)	Calibrated sink, mixed: retrieval fails from (abstention kept)	Tempering in training: keeps both
Isotropic	10	512	1,024	2,048	0.28–0.31 (0/5)	8,192–65,536 (5/5)	5/5
	20	8,192	32,768	65,536	0.95–1.06 (5/5)	— (5/5)	5/5
Clustered	10	256	256	256	0.04–0.07 (0/5)	8,192–32,768 (4/5)	1/5
	20	2,048	4,096–8,192	16,384	0.56–0.66 (2/5)	32,768 (5/5)	5/5
Heavy-tailed	10	256	256	256–512	0.19–0.23 (0/5)	4,096–32,768 (2/5)	0/5
	20	512–1,024	1,024–2,048	2,048	0.43–0.51 (0/5)	4,096–8,192 (4/5)	0/5

Table 11: Training on many lengths, five seeds each; one value where every seed agrees, otherwise the range over seeds. Fixed sink: the first length at which abstention accuracy falls below 0.9, for a sink trained at 64 keys, on lengths drawn uniformly from 64 to 1,024, and at 1,024 keys. Learned length input: the learned coefficient $\hat{\gamma}$ of $\log(N/64)$ in the null logit, and the seeds that keep both sides to 65,536 keys. Calibrated sink: sink $+ \log N$ trained on the mixed lengths; the first length at which retrieval accuracy falls below 0.9 (— if never) and the seeds that keep abstention. Tempering: the seeds in which tempering applied in training keeps both sides. A side counts as kept when its accuracy at 64 keys is at least one half and its accuracy at every tested length from 64 keys on stays at or above the lower of 0.9 and that accuracy minus 0.05, so a side can be kept at low accuracy.

When only the measured onset lies beyond the grid's last point, 65,536 keys, it passes if the predicted onset exceeds 32,768; when only the predicted onset does, it passes if the measured onset is 65,536 (Figure 13). The conditional law, with each query's own background moment, is primary; the pooled law gives the same fractions in every geometry and budget. The 16 misses are all the register's: its abstention onset with near matches at $\tau_{\max} = 20$ (cosine 0.95, four seeds; 0.80, five; 0.60, three), and its match-weight onset at cosine 0.80 and $\tau_{\max} = 10$ (four seeds).

Tempering and the heads that read the largest score. A tempered row is labelled certified within the grid if its margins, read at the largest background score of the evaluated rows, satisfy Theorem 8; explained by background concentration if the stationary law tracks its abstention to within 0.05 at every length; and empirical if it keeps both sides otherwise. None is certified. The six empirical rows are heads of the four-head geometry at $\tau_{\max} = 20$. With heavy-tailed keys the tempered rows fit no label, because their abstention fails inside the grid; the stationary law puts the mean abstention below 1/2 at 370 to 600 keys for tempering at evaluation and at 1,000 to 3,200 for tempering in training. The key check keeps both sides in 100 of its 110 rows (one per head in the four-head geometry), in every geometry except heavy-tailed keys, where it loses both, its abstention failing at the wall its extreme-value law predicts, within one grid step, in every row. The max gate keeps both sides in 95. It loses both with heavy-tailed keys, its abstention failing at the predicted wall in all nine rows in which that wall is scored, and it loses retrieval with near matches at cosine 0.60 and $\tau_{\max} = 10$. Its match weight falls below 1/2 where its law predicts, but its retrieval accuracy does not fall with it, so the registered prediction that it loses retrieval at that wall fails in 75 rows; as for softmax in the pilot (Section 6), a correct read needs only that the match's class lead the others.

Records. The extension's records (Table 16) hold every cell, the registration, the frozen reproduction gate and the assessment. `tables/tab_pilot_robustness.csv` lists, for every cell, head and side, the side statuses, the accuracy onsets, the learned τ and $b - \beta_0$, and the measured and predicted weight onsets with the regime flag (2,392 rows), and `tables/tab_mixed_length.csv` the 150 mixed-length cells, with $\hat{\gamma}$, the side statuses, the law's onsets for each cell and its controls, and the tempering labels.



Plain sigmoid: abstention is negative from $N = 128$. Row mass at 65,536 keys (seeds 0/1):
(10, 64): 689.4/688.8; (10, 512): 648.3/648.3; (20, 64): 671.4/670.7; (20, 512): 628.8/628.9.

Figure 12: Tempering keeps both weights across all four configurations; other rules need not fail within the tested lengths. Rows vary logit budget $\tau_{\max}$ and training length N_{train}; columns show absent-query abstention $a = 1 - \sum_j w_j$, present-query match weight w and retrieval accuracy. Colour says how a rule's null scales with the number of keys N: fixed (orange), raised by $\ln N$ (teal) or tempered together with every score (violet); softmax (blue) has no null. Shape says the mechanism: a sink (circle), a register (square) or a sigmoid (diamond); softmax is a triangle. Seven rules, two seeds, 13 lengths (16–65,536 keys), 1,024 queries per length and seed. A filled mark and a ring are seeds 0 and 1; dashed lines are the law's prediction for each seed's trained head. Plain sigmoid's 80 negative abstention values are off scale. At $(20, 512)$, both calibrated rules retain accuracy one, but sigmoid $-\ln N$ has match weight below one half. Tempering is empirical, not certified; $\lambda = \ln N / \ln N_{\text{train}}$ applies below training too. Absent queries have N background keys, present queries $N - 1$; the theorem uses N in both.

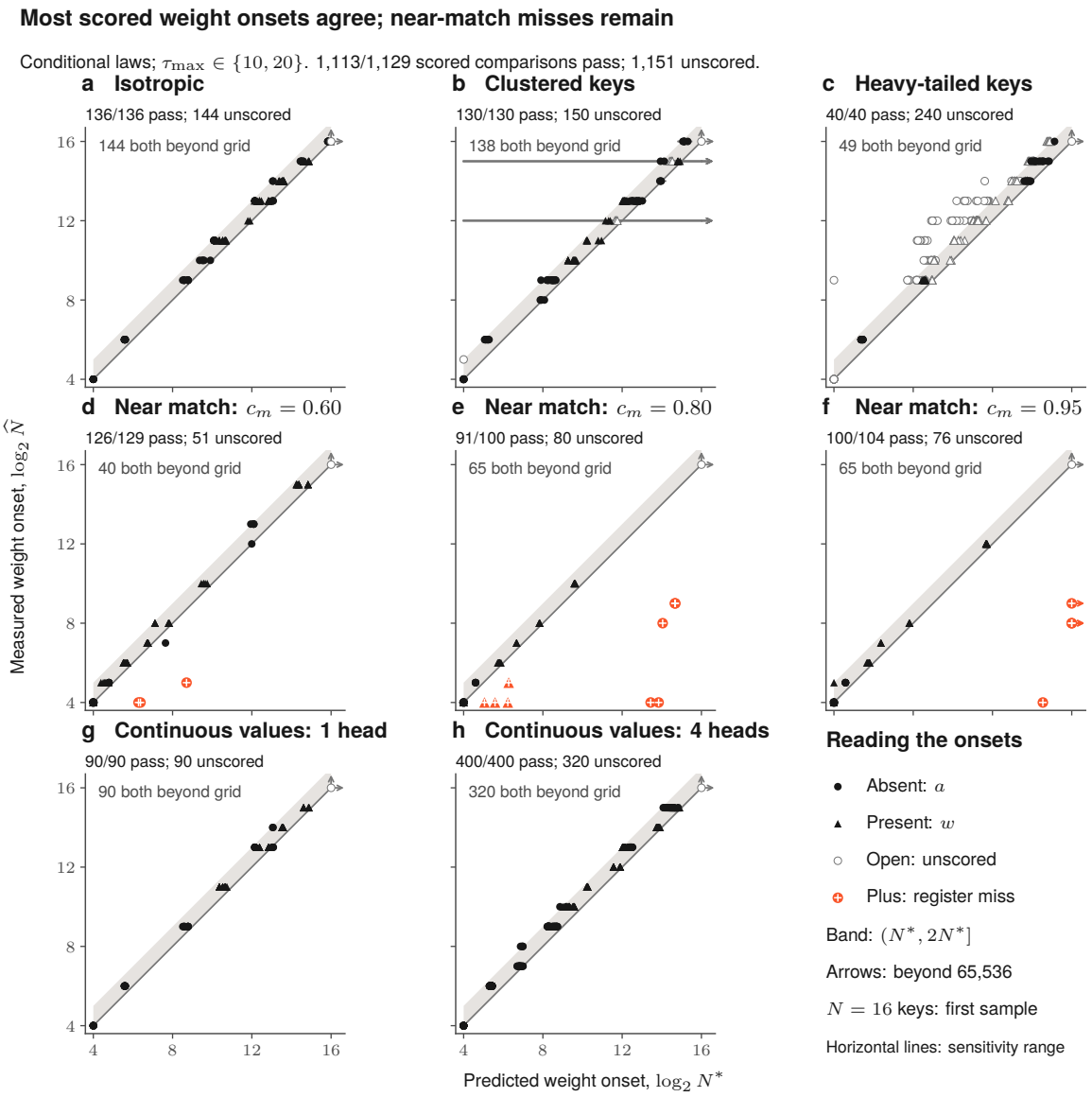


Figure 13: Most scored weight onsets agree with the law (1,113/1,129); all 16 misses are near-match register cases. (a–h) Onsets below one half, predicted N^* versus first measured $\widehat{N}$: absent-query abstention a (circles), present-query match weight w (triangles), eight geometries. Filled marks are scored; hollow ones unscored, including outside-regime comparisons. All 2,280 comparisons appear: 1,366 pairs in range, three measured failures predicted beyond it, and 911 with neither onset in range (untestable, counted corner arrows). Arrows mean beyond 65,536 keys; $N = 16$ keys is the first sample. Comparisons can share a location. The band shows one doubling; acceptance requires $|\log_2(\widehat{N}/\widetilde{N})| \leq 1$, with $\widetilde{N}$ the first tested length $\geq N^*$, plus beyond-range rules (Appendix B.3). Horizontal brackets in (b) show onset sensitivity to poorly resolved extreme background scores; scoring requires the comparison's decision to hold at both ends. Neither brackets nor band are confidence intervals. Every head's two weights are included: the 927-cell extension, logit budgets $\tau_{\max} = 10, 20$, five seeds per setting. Only conditional laws are shown, without fitting; comparisons are not independent trials (Table 16). At logit budget 20, near-match $c_m = 0.80, 0.95$ pass only $20/25, 25/29$, below the registered 90% criterion. Mixed/partial: calibrated abstention drifts with heavy-tailed keys; tempering fails with clustered keys. Finite-grid retention does not certify tempering.

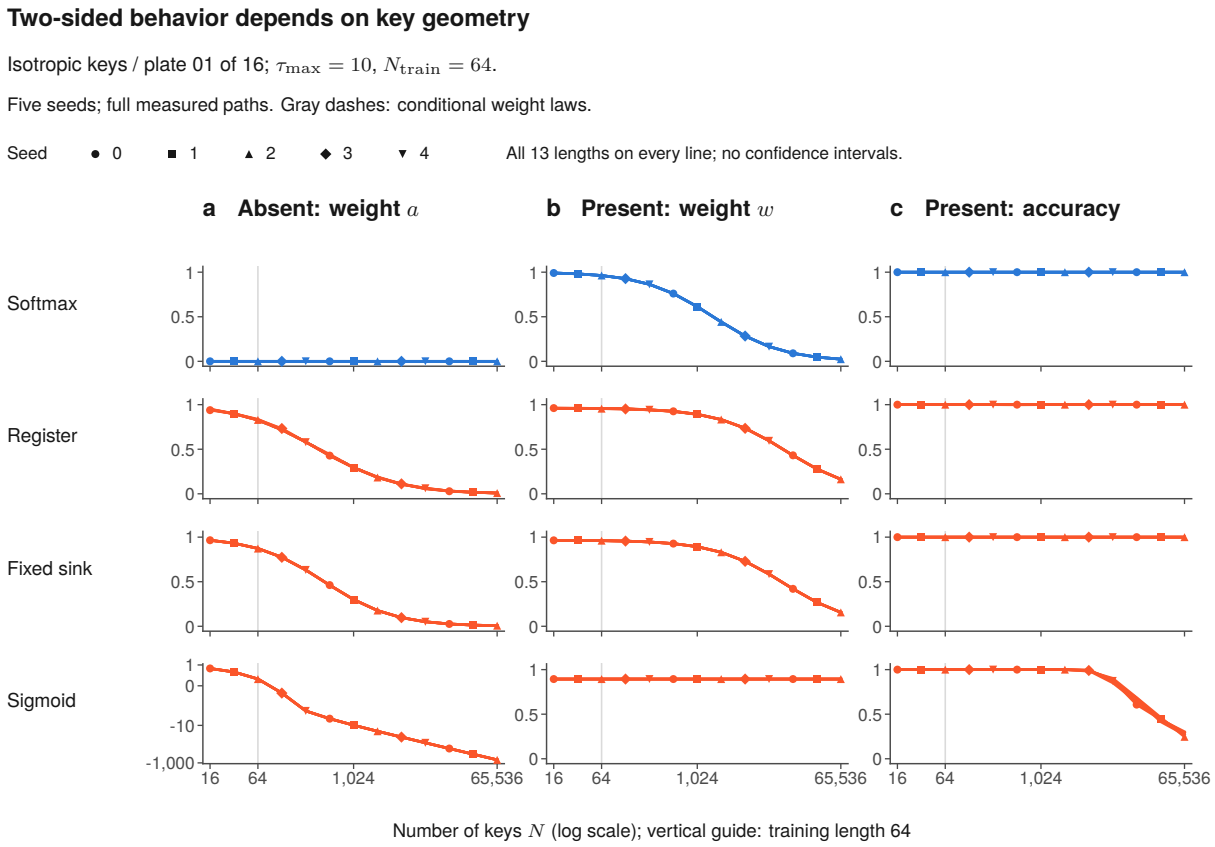


Figure 14: The two-sided behavior depends on key geometry; weight readouts and retrieval accuracy remain distinct. This sixteen-plate atlas shows every geometry cell at logit budget $\tau_{\max} = 10$ and $N_{\text{train}} = 64$: eight geometries, nine rules, five seeds (0–4), and thirteen lengths $N = 16, 32, \ldots, 65{,}536$. Plate A contains softmax, register, fixed sink and plain sigmoid; plate B contains the two calibrated rules, log-N tempering and the two maximum-statistic rules. The columns show absent-query weight $a = 1 - m$, present-query match weight w, and present-query accuracy. Colour identifies fixed nulls (orange), length-calibrated nulls (teal), tempering (violet), and softmax without a null (blue); the two maximum-statistic rules use neutral ink. Solid paths connect all 13 samples for each seed; shapes identify seeds without shifting observations. Gray dashes draw every conditional weight law, with no law on accuracy: one head per seed, or all four heads per seed for the final geometry (no averaged law). Its measured weights are the overall readouts. All 4,680 curve samples (14,040 measured values) and 6,435 head-law pairs are shown. Plain sigmoid's absent-weight axis uses a signed logarithmic scale, linear on $[-1, 1]$; calibrated sigmoid uses a linear scale including its negative values. All 425 negative measurements are visible and are not abstention probabilities. The vertical guide is the training length. Individual seed paths show the spread; no confidence interval or independent-head interpretation is added. The extension's outcome is mixed/partial: heavy-tailed keys violate calibrated-abstention retention and clustered keys violate tempered retention. Near-match register qualifications at logit budget 20 and mixed-length training belong to the complete assessment, outside this logit-budget-10 view. Weight-onset scoring and its unscored cases are reported separately in Figure 13; the registered per-geometry rule is unchanged. Tempering is certified nowhere. Maximum-statistic rules are outside the sum-rule refutation class. See Table 16 for provenance.

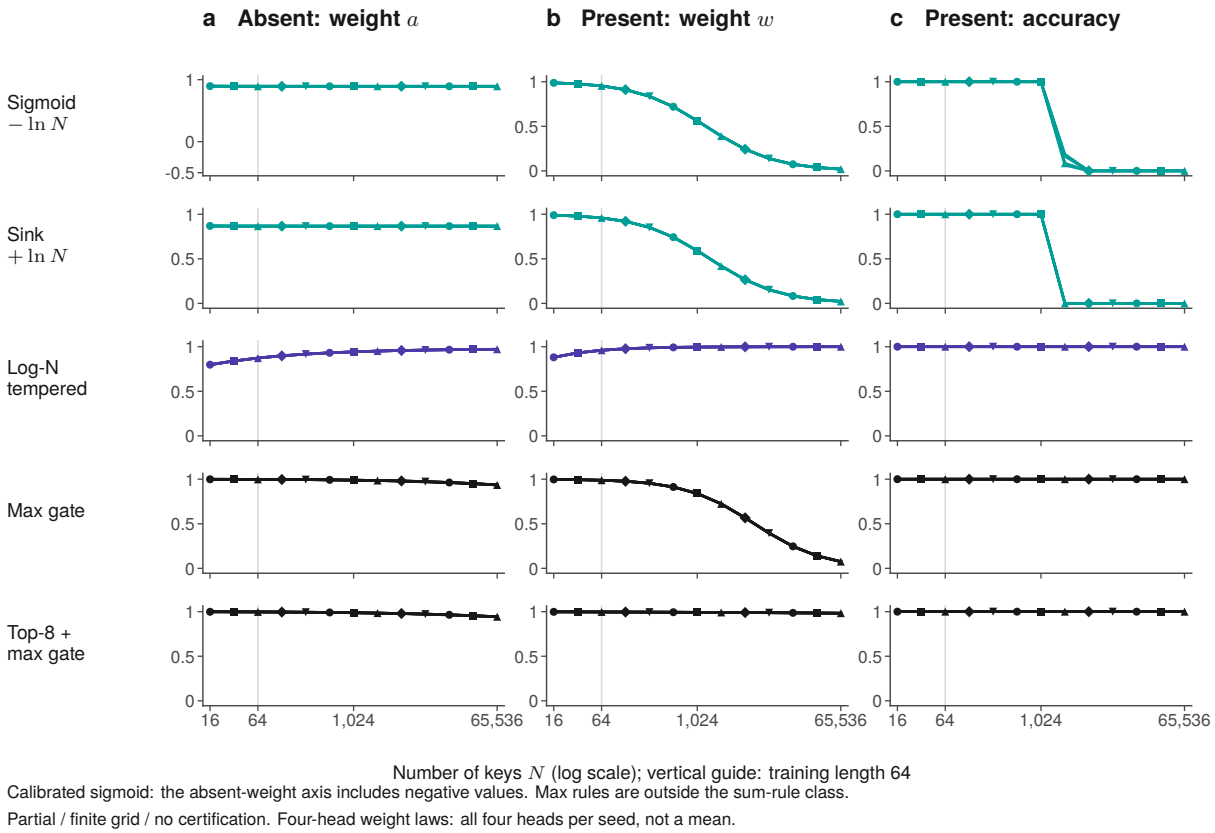


Figure 14 (continued), plate 02/16: Isotropic keys, part B. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

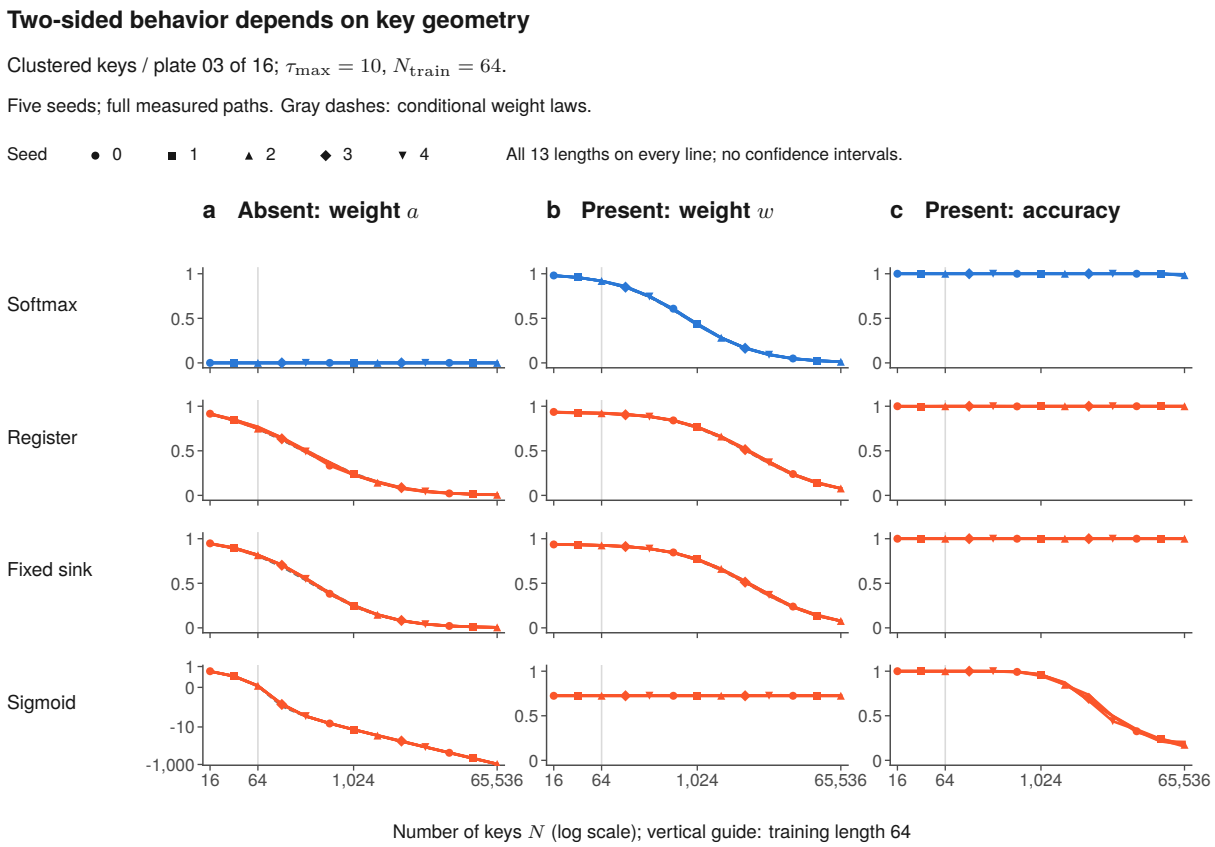


Figure 14 (continued), plate 03/16: Clustered keys, part A. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

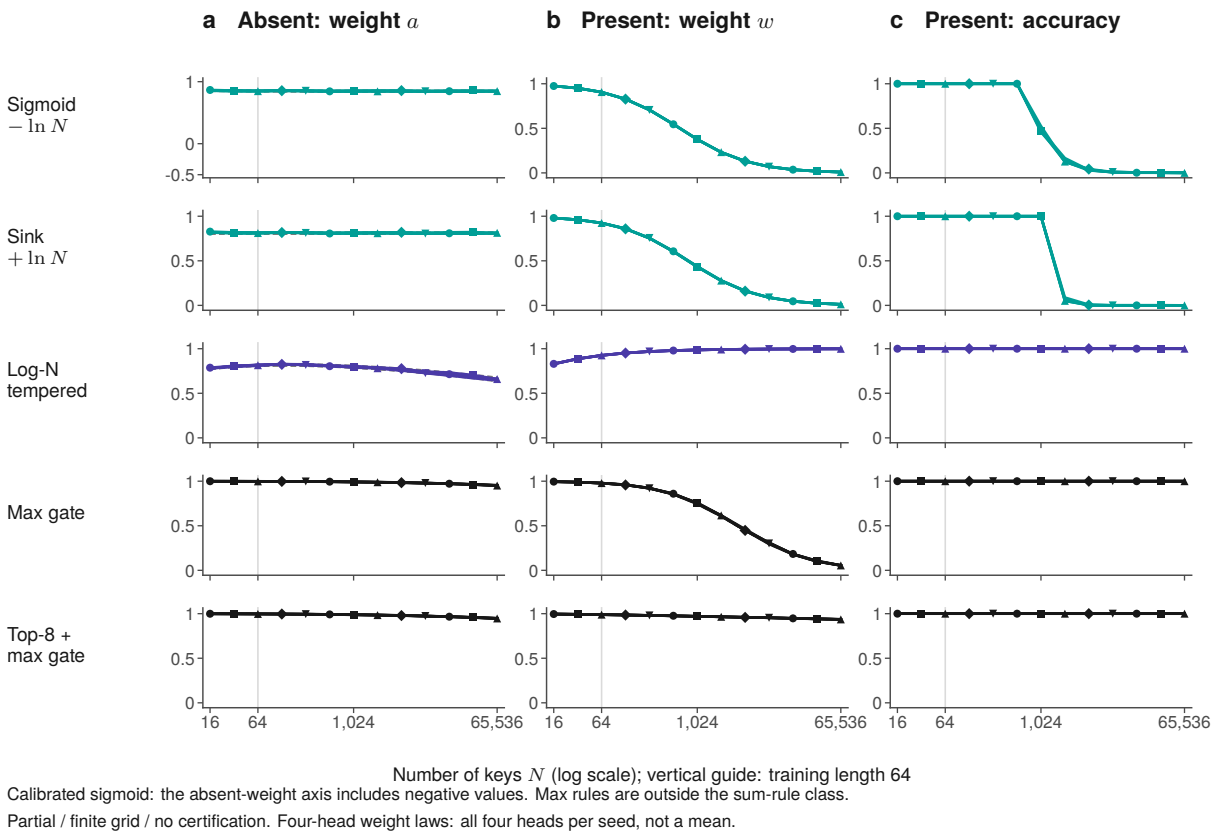


Figure 14 (continued), plate 04/16: Clustered keys, part B. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

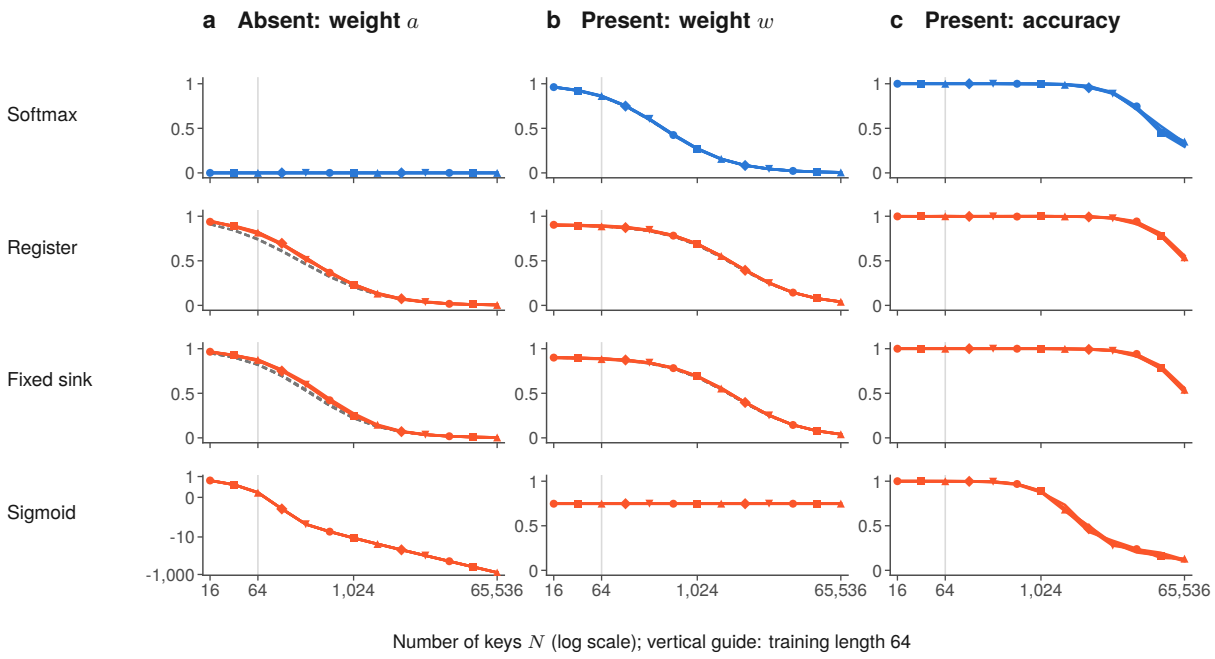


Figure 14 (continued), plate 05/16: Heavy-tailed keys, part A. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

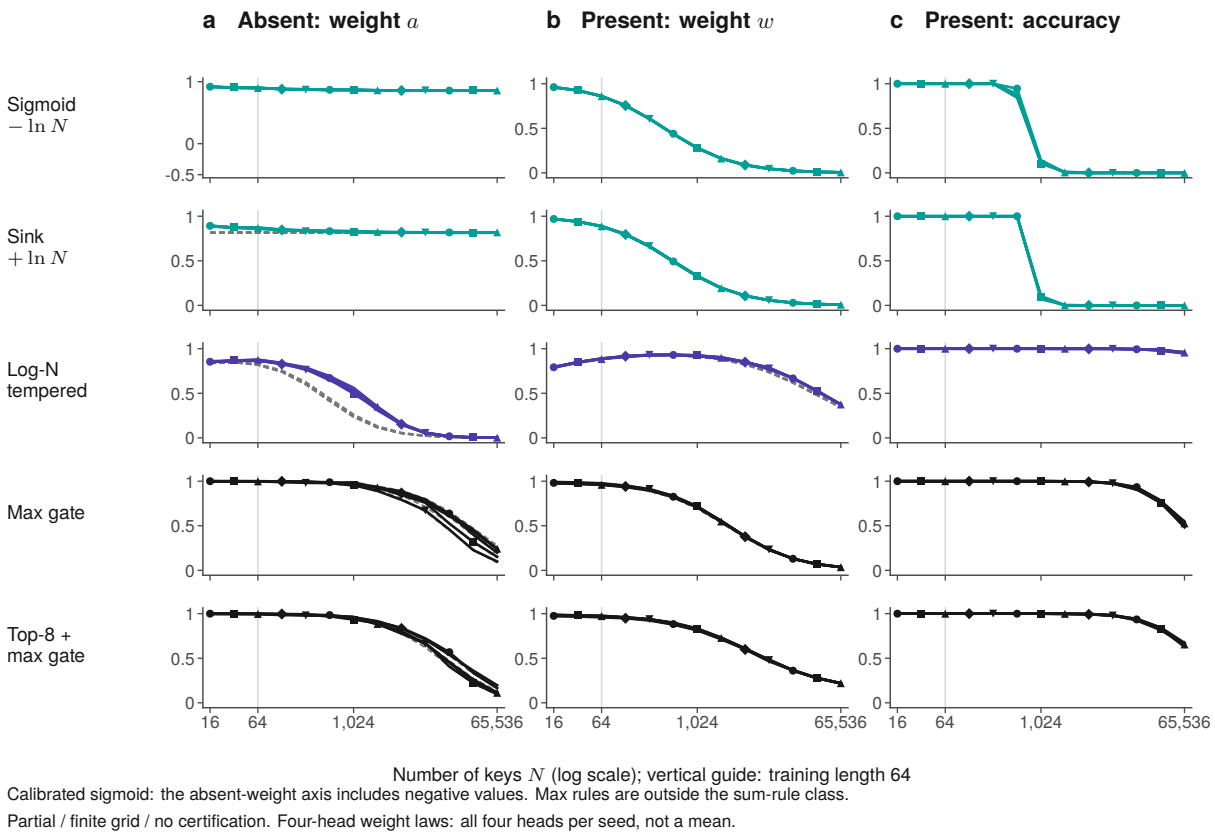


Figure 14 (continued), plate 06/16: Heavy-tailed keys, part B. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

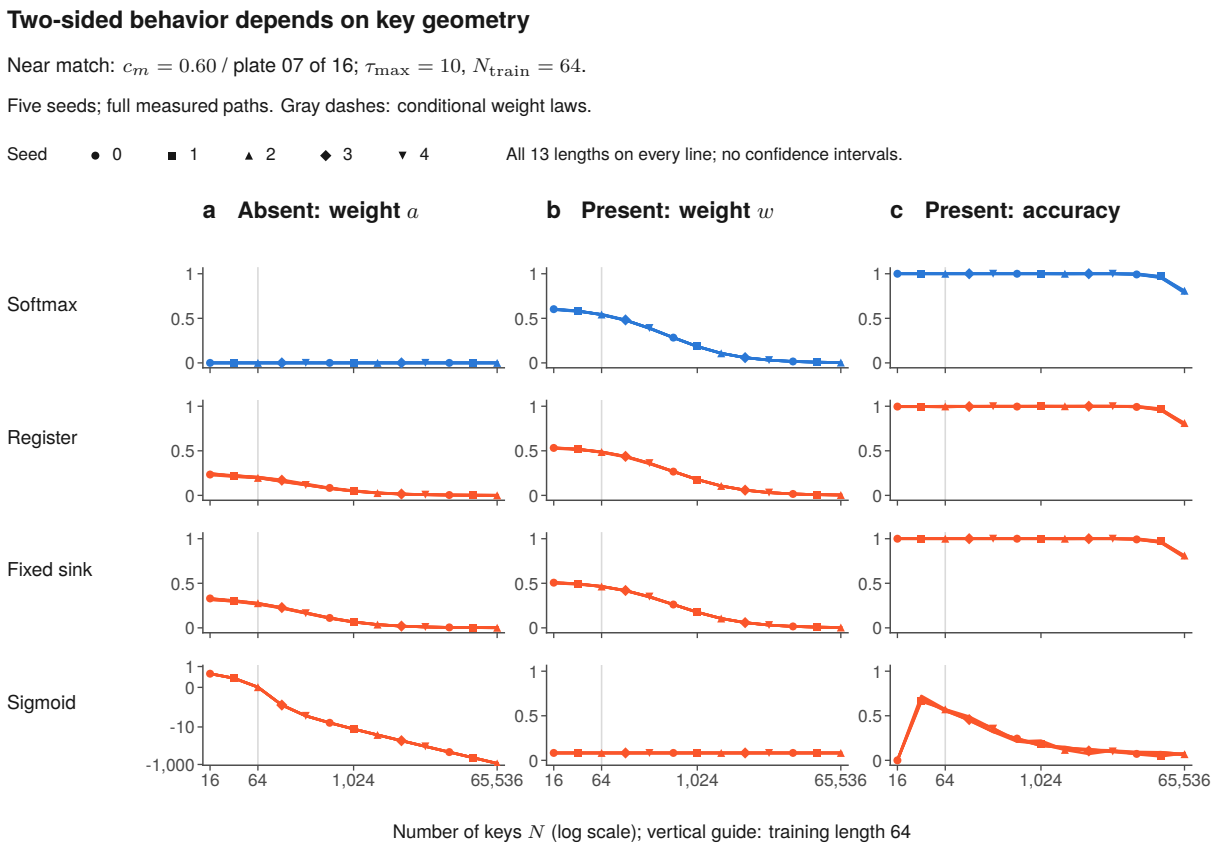


Figure 14 (continued), plate 07/16: Near matches at $c_m = 0.60$, part A. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

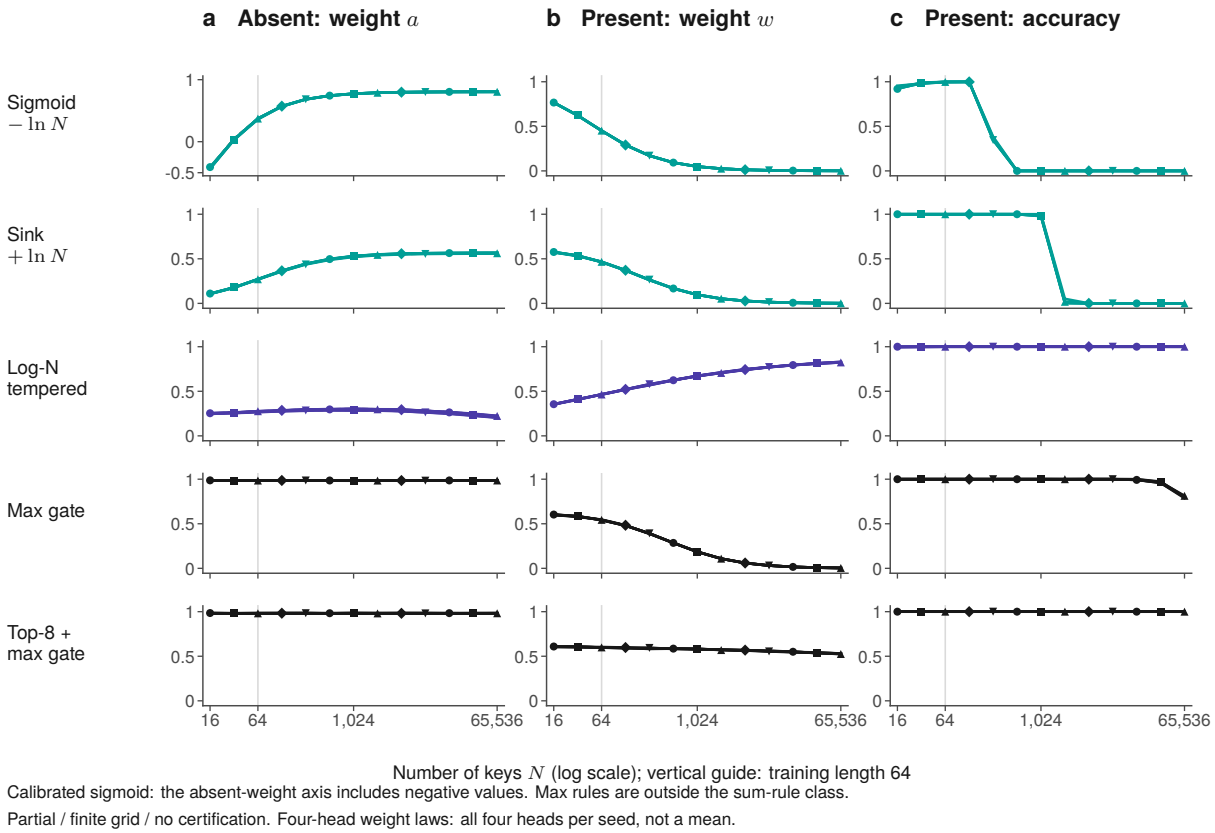


Figure 14 (continued), plate 08/16: Near matches at $c_m = 0.60$, part B. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

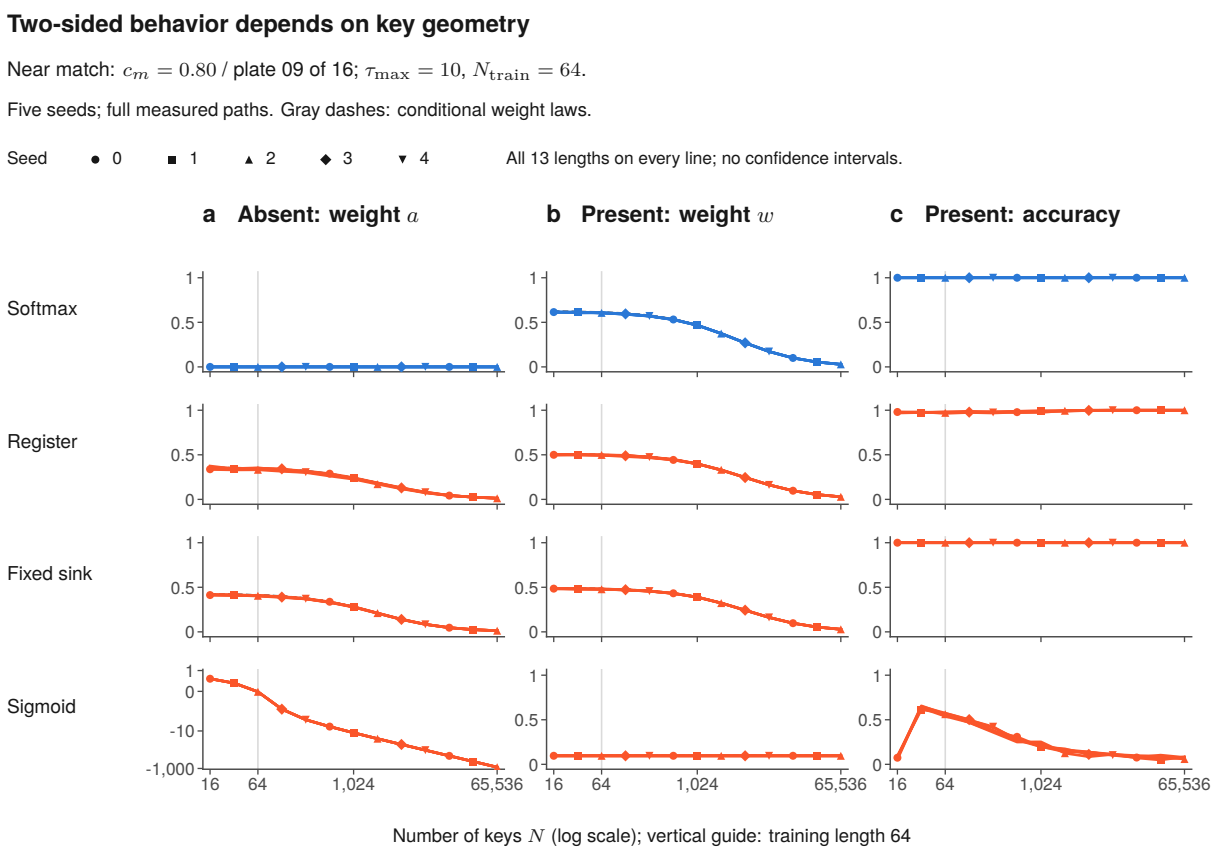


Figure 14 (continued), plate 09/16: Near matches at $c_m = 0.80$, part A. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

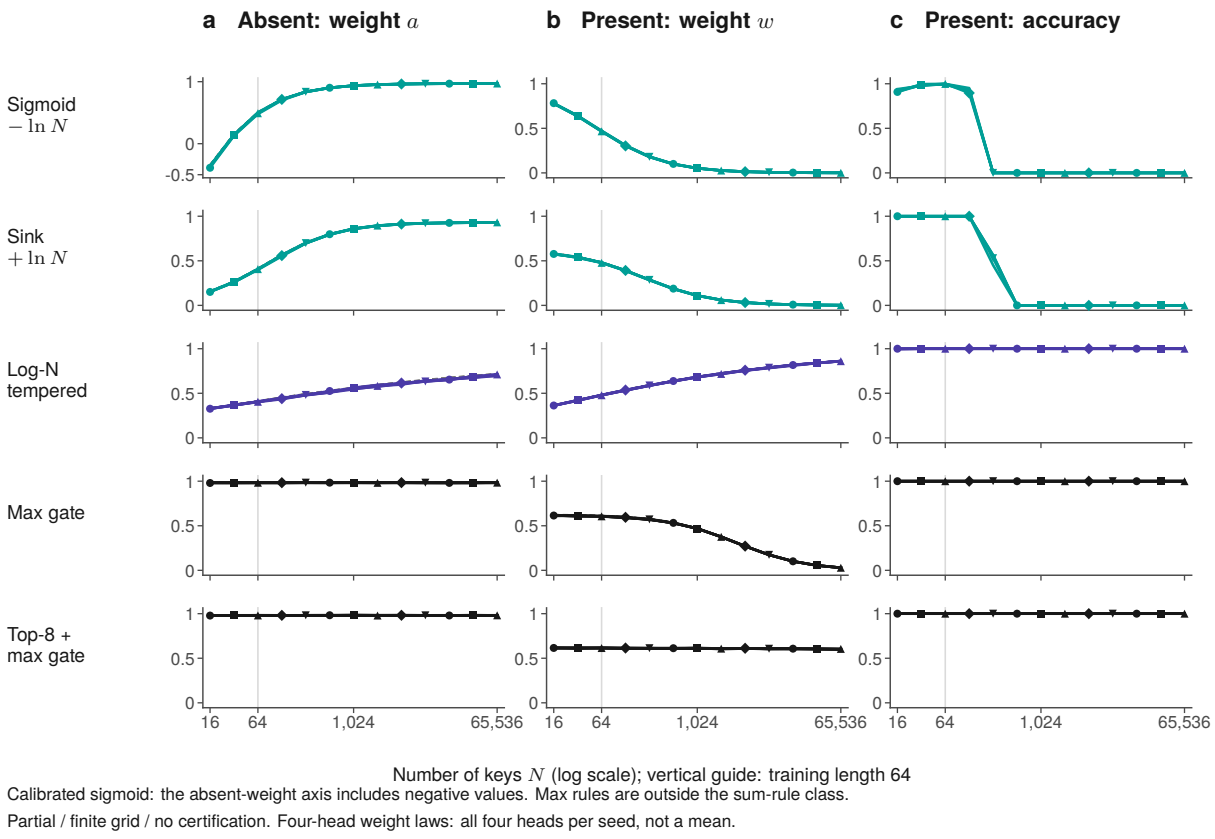


Figure 14 (continued), plate 10/16: Near matches at $c_m = 0.80$, part B. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

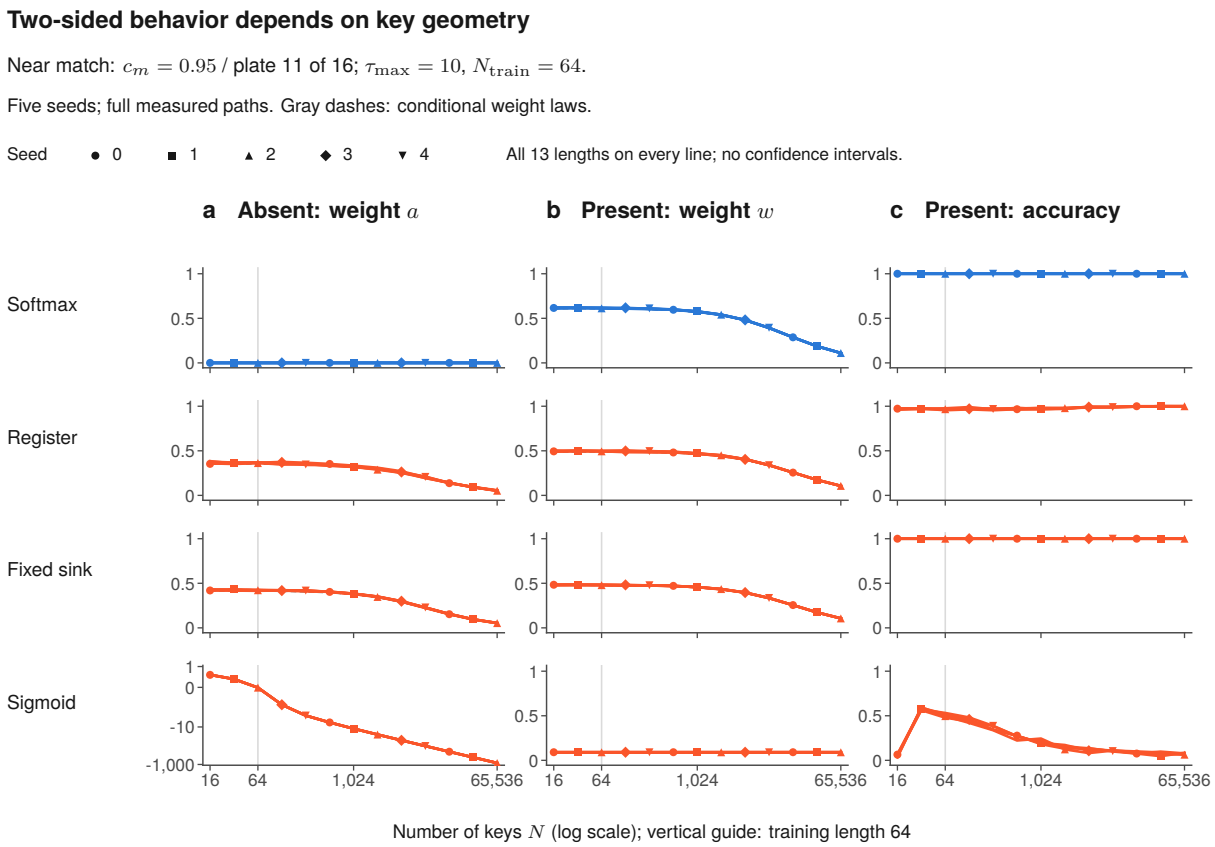


Figure 14 (continued), plate 11/16: Near matches at $c_m = 0.95$, part A. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

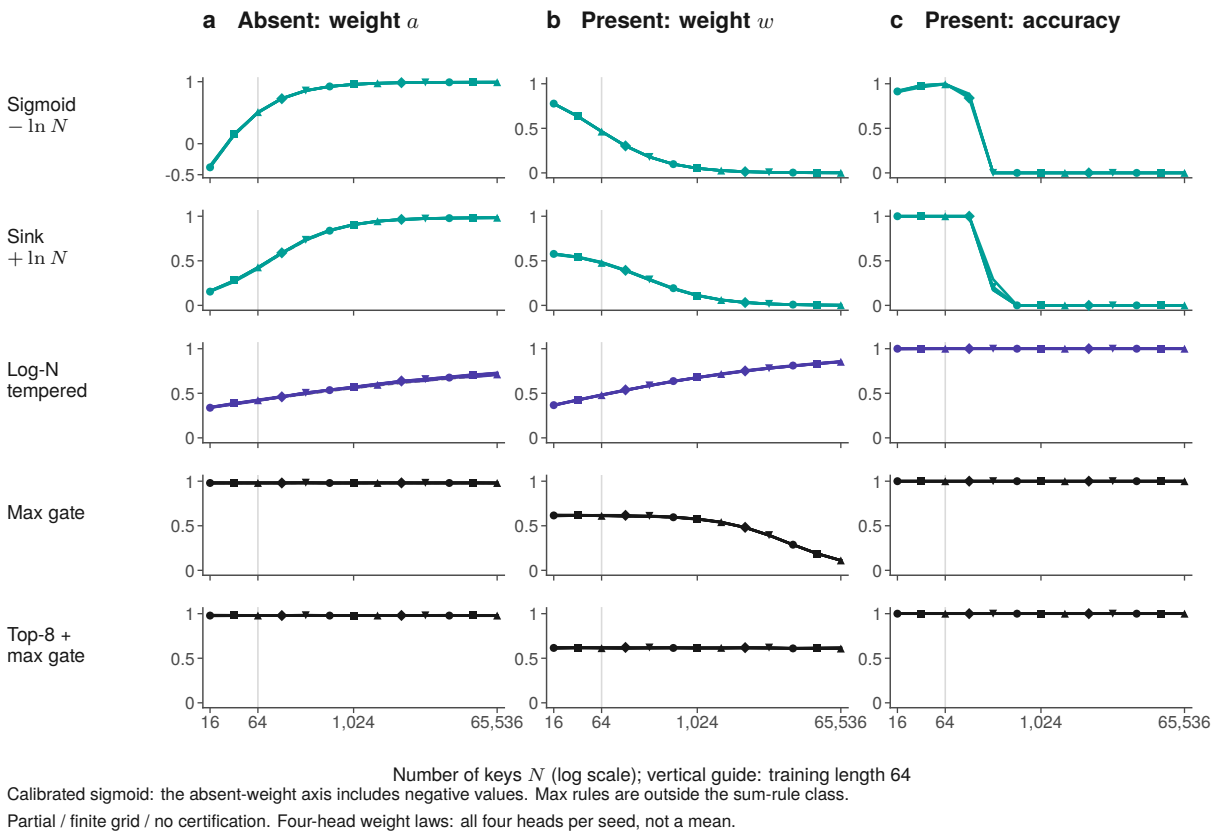


Figure 14 (continued), plate 12/16: Near matches at $c_m = 0.95$, part B. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

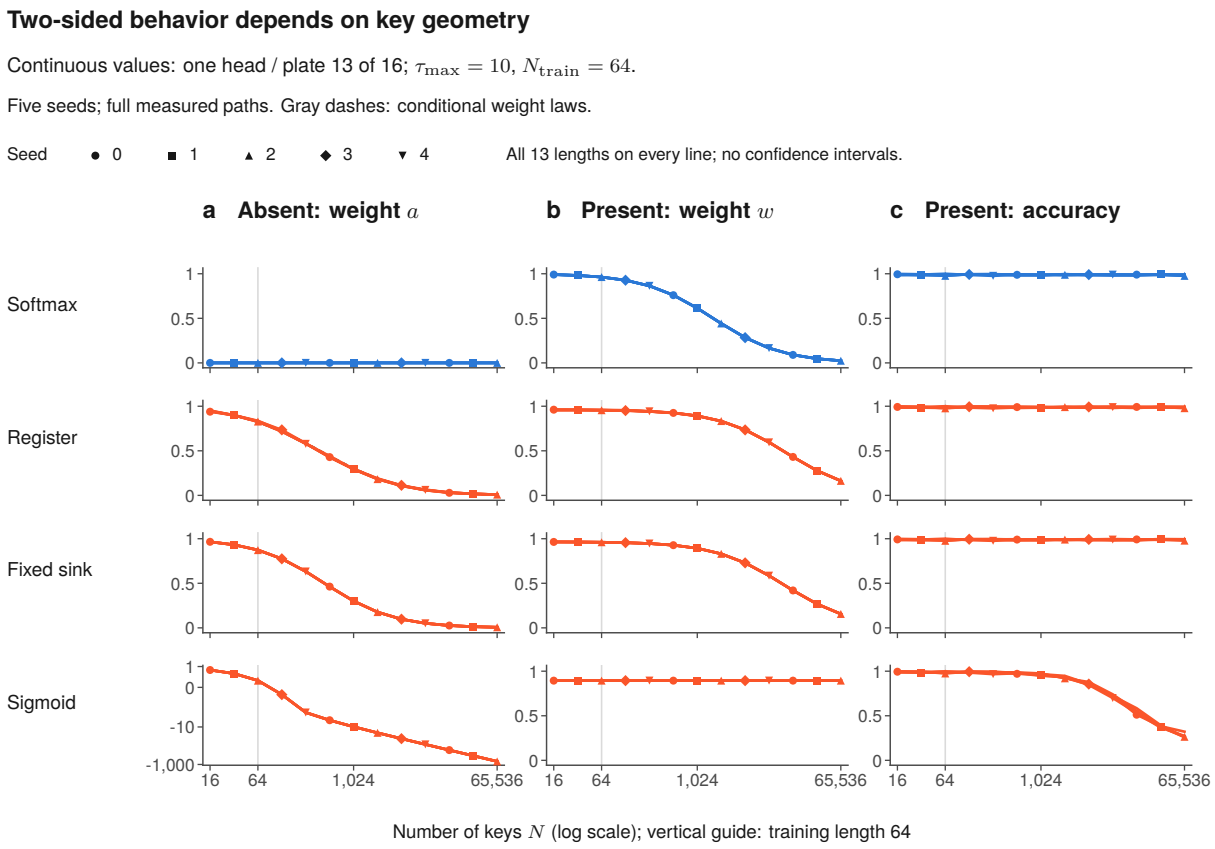


Figure 14 (continued), plate 13/16: Continuous values, one head, part A. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

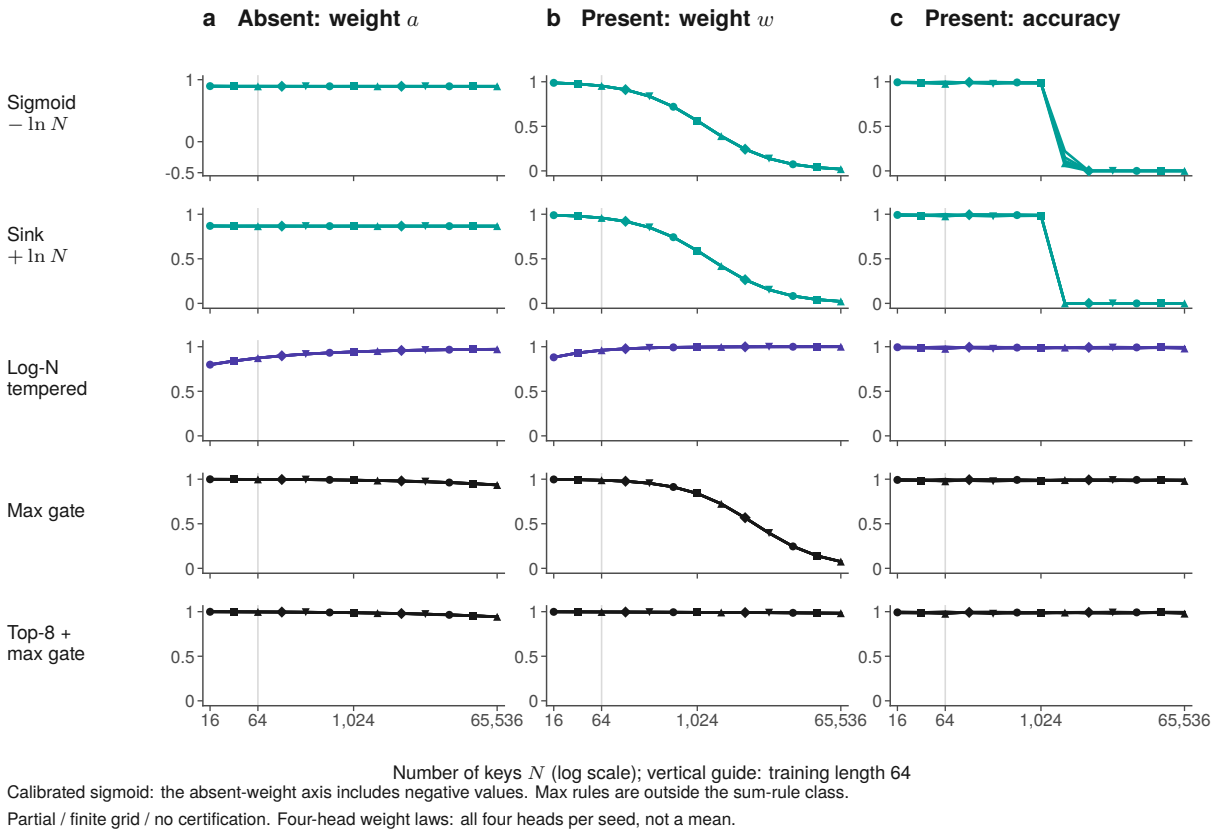


Figure 14 (continued), plate 14/16: Continuous values, one head, part B. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

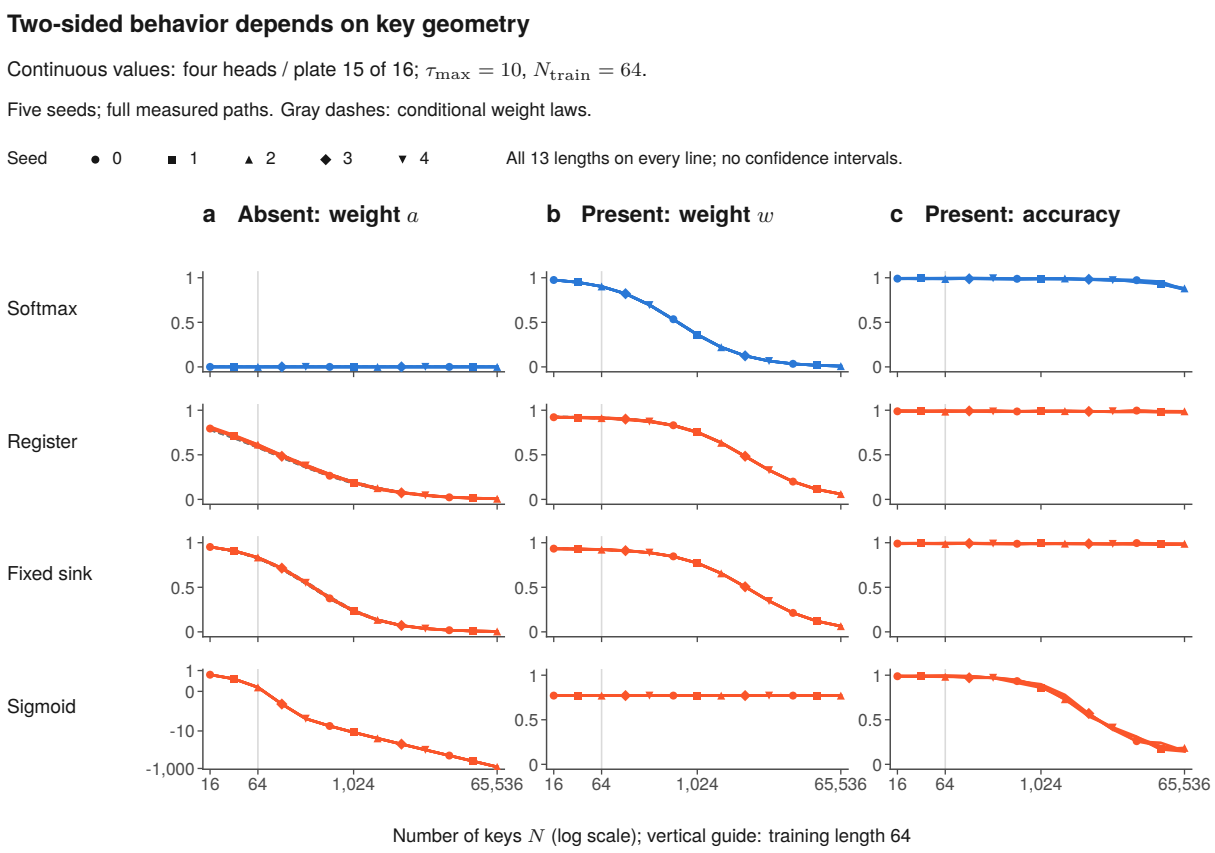


Figure 14 (continued), plate 15/16: Continuous values, four heads, part A. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

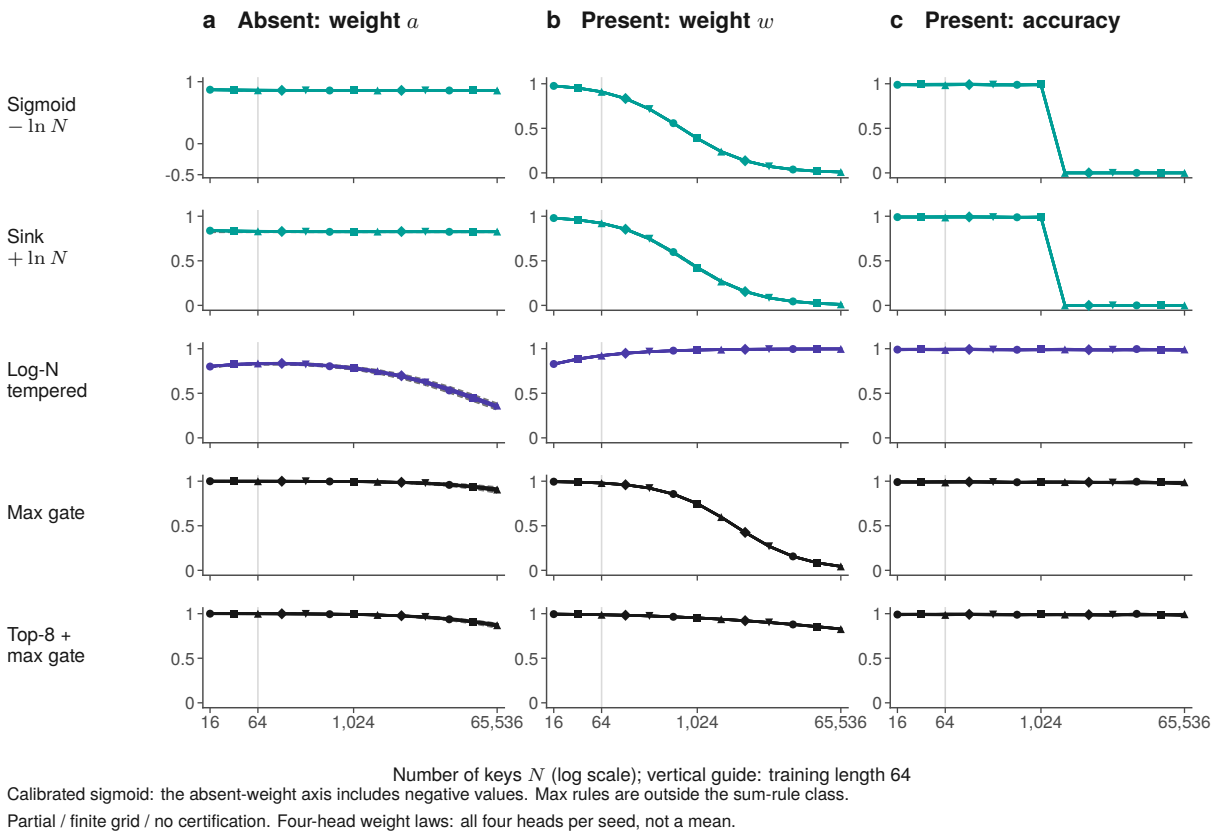


Figure 14 (continued), plate 16/16: Continuous values, four heads, part B. Columns and encodings are unchanged: absent weight, match weight, and present-query accuracy; five seeds and all thirteen sampled lengths. Gray dashes show saved conditional weight laws, with all four head laws per seed in the four-head geometry. Every plate is descriptive finite-grid evidence; tempering is not certified.

C The prespecified gpt-oss protocol

Model and forward pass.

- **Model:** `openai/gpt-oss-20b` at revision `6cee5e81...`.
- **Precision:** the attention weights are stored in bfloat16. The mixture-of-experts weights are dequantized from MXFP4 to bfloat16, which is exact.
- **Run settings:** batch size 1 with no padding, and 4,096 tokens per document.
- **The probe:** a custom attention interface calls the model's own eager attention unchanged for the output, so the forward pass is bit-identical to eager. It computes the measurement on the side from the same queries, keys, mask and sink logits.
- **The log-odds:** $y_h(i) = b_h - \mathrm{LSE}_{j \in K(i)}\, s_{ij}$ is computed in float32 from the layer's bfloat16 queries and keys, with the visible set $K(i)$ read from the mask the layer actually receives.
- **Environment:** the runs used one NVIDIA H100 80GB with torch 2.6.0, transformers 5.11.0 and CUDA 12.2. TF32 was off and float32 matrix products were at the highest precision.

Documents and conditions.

- **Books:** the first 32 PG-19 test-split books, in canonical order, whose body has at least 8,192 tokens.
- **Code:** the first 32 CPython standard-library files, in sorted path order, with at least 8,192 tokens, excluding `test/` and `idlelib/`.
- **Window and deduplication:** each document is read as tokens $[0, 4096)$ of its body, with no special prefix token. A document whose first 1,024 tokens share more than half their 8-grams with another's is dropped.
- **The shuffled condition:** each window's tokens are permuted with the generator seeded at 10,000 plus the document index. The mask and the key counts are unchanged.
- **The manifest:** the document list is frozen in the analysis bundle (SHA-256 `cd0a903d...`).

Statistics.

- **Sink-using heads:** a head qualifies if its mean sink probability $\sigma(y_h)$ over the natural documents and positions 8–64 exceeds 0.05.
- **Curves and slopes:** for each head, the document-mean curve $\bar{y}_h(i)$ is fitted by ordinary least squares on $\log(i+1)$ over a window of positions. The fitted slope equals the mean of the per-document slopes.
- **The legs:** each leg is the median slope over the sink-using heads of one layer type: dense layers at positions $[256, 4096)$, holding at a median of at most -0.5; banded layers at $[256, 4096)$, holding at an absolute median below 0.2; and banded layers while the window fills, at $[8, 128)$, holding at a median of at most -0.5 (Section 9).
- **Intervals:** 2,000 bootstrap resamples of documents, stratified by corpus, recompute the curves, the head set and the medians (seed 0).
- **The cross-fitted check:** selection overlaps the third leg's fit window, so heads are also selected on the odd documents of each corpus and fitted on the even ones, and vice versa. A third leg whose sign differs from the primary statistic's in either direction is labelled selection-sensitive; the label gates nothing.

Gates and the decision rule. Each production cell must pass:

- G1: exact key counts in every row of every layer;
- G2: one probe call per layer;
- G3: a bit-identical forward pass on the first document;
- G4: bfloat16 fidelity, a median $|\sigma(y) - (1 - \sum_j w_j)|$ of at most $\tau_{\mathrm{fid}} = 10^{-3}$ on rows with $\sigma(y) \in [0.02, 0.98]$, where τ_{fid} was fixed by a pilot before the freeze;
- G6: finite sink logits, read from the configuration's 12 dense and 12 banded layers.

The assessment then applies the corpus control G5: the mean natural next-token loss must lie more than 1 nat below the shuffled loss, in each corpus. Then it takes the branch:

- confirm, if all three legs hold;
- compensated, if the dense leg fails and the two banded legs hold;
- inconclusive, otherwise, or if a layer type has fewer than 16 sink-using heads.

Point estimates decide the branch, and the outcome text says "marginal" when an interval crosses its threshold. Path-law quantities for a companion paper were recorded under the same freeze. They play no part in the branch and are not read here.

The freeze. The analysis bundle holds:

- the prespecification file;
- frozen copies of the analysis and probe functions;
- the document manifest;
- the protocol text.

Its SHA-256 is `bd82a560...`, and it was frozen at 17:04 UTC on 30 September 2026. Every production cell checked its members against the files on disk before running, and the first capture began at 17:12 UTC. The bundle was not publicly deposited before the run, so the analysis is prespecified, not publicly registered.

Deviations.

- **The validator.** The original pipeline's pilot ran on one H100, but its validator failed: integer layer keys sort differently after a JSON round trip.
- **The recovery path.** The production captures therefore ran through a recovery path with separate entry points. It binds the untouched pilot record, the original producer sources and the recovery sources, and the scientific contract did not change.
- **The specification's hash.** It differs from the draft's by a 391-byte dated staging note. Removing that note reconstructs the draft's hash exactly.
- **The fidelity tolerance.** The tolerance $\tau_{\text{fid}} = 10^{-3}$ was set from the pilot's median error of 2.5×10^{-4} and frozen before the main captures.

Outcome. All implementation gates pass in all four cells. G5 passes for code (gap 3.87 nats) and fails for books (0.672 nats), and the legs match neither decisive branch: the banded leg beyond the window holds, while the dense leg and the banded leg as the window fills are flatter than -0.5. So the branch is inconclusive; descriptively, the test leans toward compensation without settling it (Section 9, Table 6). The assessment's SHA-256 is `5b06be82...`.

D Experimental details and records

This appendix gives the training and evaluation details of the models trained for Sections 7 and 8, the records behind the trained-model tables, the protocol of the production-ViT census, and, in Table 16, the folder of the package's `data/` directory that holds each experiment's records.

D.1 The trained models

The 125M language models. Every rule shares one decoder: 12 pre-norm blocks of width 768 with 12 heads, rotary position embeddings (base 10,000) on queries and keys, a GELU MLP of ratio 4, and the GPT-2 vocabulary padded to 50,304 with tied embeddings. Softmax has 123.7M parameters, and no rule adds more than 0.1% to them. Each model is trained on FineWeb-Edu for 9,537 steps of 256 sequences of 1,024 tokens (2.5B tokens) with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$; weight decay 0.1, not applied to biases, norms, embeddings, the registers, the weights of the gate on each value or the rules' scalar parameters), a peak learning rate of 1.2×10^{-3} reached after a 2% warmup and decayed by a cosine to a tenth of it, gradient clipping at 1.0, and bfloat16 arithmetic in the dense layers with float32 attention. A 2,000-step pilot of five rules chose the learning rate from 0.5, 1 and 2 times 6×10^{-4}, and every rule uses it. Three seeds per rule give 33 checkpoints. The perplexity ratio D is the perplexity of the first L tokens of 95 PG-19 test books divided by that of their first 1,024, for $L = 1{,}024$ to 8,192. YaRN interpolates the rotary frequencies by parts

with scale $L/1{,}024$ and multiplies every query–key score, the registers' included, by $(0.1\ln(L/1{,}024)+1)^2$; it does not scale a virtual sink's logit, except in a tempered rule, whose scores and sink logit both carry both factors. Passkey retrieval hides a five-digit key at depths 0.1, 0.5 and 0.9 of PG-19 filler, 50 trials per length and depth, and scores the key's log-probability against nine random keys (chance 0.1). The probe places 16 facts of the form "The code of *name* is *v*." in PG-19 filler and asks for one that is present or one that is absent, 64 trials per length and condition. Retrieval heads are fixed before scoring, on 64 identification trials at 1,024 tokens: the heads whose mean attention from the query to the matching fact reaches 0.3, or the three with the most if fewer qualify. Sinks are read on 64 FineWeb-Edu validation sequences at 1,024 tokens and on 32 PG-19 windows at 8,192 tokens, and massive activations on the 64 sequences. The re-evaluation that separates interpolation from temperature reuses the checkpoints of softmax, the learned sink, the output gate and the key check on the first 24 of the books, at 2,048 to 8,192 tokens, applying YaRN's interpolation and its temperature separately, alone on all four rules and with the schedule of Theorem 8 on the first three.

The ViT subset. ViT-S/16 (12 blocks of width 384 with six heads, a CLS token and learned absolute position embeddings; 21.7M parameters for softmax) is trained on the 126,689 ImageNet-100 training images at 224 px with five rules: softmax, four value-null registers, a learned sink, an output gate, and the gate with a learned sink. Training follows the DeiT recipe (Touvron et al., 2021) for 300 epochs at batch size 256: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.05), a peak learning rate of 2.5×10^{-4} after five warmup epochs and a cosine decay to 10^{-5}, mixup 0.8, CutMix 1.0, label smoothing 0.1, random erasing 0.25, stochastic depth 0.1, gradient clipping at 1.0 and bfloat16 arithmetic with float32 attention. An exponential moving average of the weights (decay 0.9999) is the evaluated model. A 50-epoch pilot, scored on 50 images per class held out from the training images, chose the learning rate from 0.5, 1 and 2 times 5×10^{-4}, and all five rules preferred the same one. Three seeds per rule give 15 checkpoints. Top-1 accuracy uses the 5,000 validation images. Higher resolutions resize the 224 px evaluation crop bicubically and interpolate the position embeddings bicubically; tempering counts the patches and CLS in N and leaves registers out. Sink statistics use 64 validation images, the first of each of the first 64 classes. The prespecified confirmation adds 896 px and softmax tempered the same way, on the same images. The attention-row re-evaluation uses the nine checkpoints of softmax, the learned sink and the gate with a learned sink, the CLS rows of the first 1,000 validation images at 224 to 896 px, and two input routes: the upsampled crops and the original images resized to each resolution. Its nine records missed the registered check that recomputes their summary exactly (the fitted values differ between recomputations by at most 2.4×10^{-15}), and an amendment to the registration admits them at a tolerance of 10^{-12} with every measurement and verdict unchanged.

The byte-level models. The byte-level language models trained across lengths and positional schemes in Section 7 share one design: a byte embedding with a causal depthwise convolution over the last four bytes, four transformer blocks of width 128 with four heads, and a final norm, 594,304 parameters before any null. They are trained on the raw WikiText-2 training text with AdamW (peak learning rate 2×10^{-3} in a one-cycle schedule with 10% warmup; weight decay 0.05, not applied to null logits or the output gate's bias) for 1,500 steps of 4,096 bytes, with gradient clipping at 1.0 and five seeds per setting. The models that ask whether training holds abstention fixed have no positional encoding, are trained at 64, 128, 256, 512 and 1,024 bytes, and are evaluated on 256 validation windows at their training length and at 128, 512 and 2,048 bytes. The models that separate positions from the null are trained at 256 bytes with no positions, rotary positions (base 10,000) or rotary positions on a quarter of each head, and with no null, a learned null, a learned null plus $\ln n$, or an output gate, and are evaluated on 128 validation windows of 2,048 bytes, with attention rows read at positions 63, 127, 255, 511, 1,023 and 2,047 (15 rows at each). The fifteen models with a learned null plus $\ln n$ missed the registered check on the recomputed null mass (tolerance 10^{-6}; errors of 1.7 to 2.0×10^{-6}), and an amendment to the registration admits them with every recorded value unchanged.

D.2 Trained-model records and the production-ViT census

Trained-model records. Table 12 gives every seed's value behind the comparisons of Table 3 that have fewer than five seeds. Table 13 gives the prespecified within-head test of sink mass against the contraction defect beyond the mixing term (Section 7). Its columns run over each model's heads, a head's sink mass being its mean first-token weight over its content queries: m is the median sink mass; S_0 is the Spearman

correlation of sink mass with the part of the defect that the sink masses fix by themselves, near one by construction; S_1 is the partial Spearman correlation of sink mass with the total defect given the mixing term, with the value that construction alone gives, S_1^0, in parentheses; S_2 is the Spearman correlation of sink mass with the mixing term; $\widetilde{\psi}$, the decision statistic, is the median over eligible heads of the within-head Spearman correlation between successive changes, over query positions, in a row's null factor $s(2-s)$ and in its routing concentration above uniform; S_4 is the Spearman correlation of sink mass with the earlier census's regularized defect; and the last column is the share of heads whose content operator is not a contraction.

Change	Readout	Arm	Seed 0	Seed 1	Seed 2
Sigmoid, depth 12	content-sink mass	softmax	0.01893	0.0485	0.01132
		sigmoid	0.00144	0.004878	0
	layers with a sink	softmax	0.6667	0.6667	0.25
		sigmoid	0.3333	0.08333	0.25
	row mass	softmax	1	1	1
		sigmoid	0.3494	0.3258	0.3493
	massive activation	softmax	39.61	36.64	28.05
		sigmoid	43.29	68.85	49.92
	test accuracy (%)	softmax	55.54	55.92	56.18
		sigmoid	53.17	52.96	53.9
Sigmoid, depth 24	content-sink mass	softmax	0.03982	0.02232	0.03661
		sigmoid	0.002755	0.000698	0.0008294
	layers with a sink	softmax	0.625	0.5833	0.75
		sigmoid	0.08333	0.08333	0.04167
	row mass	softmax	1	1	1
		sigmoid	0.3211	0.3185	0.302
	massive activation	softmax	80.04	51.68	50.9
		sigmoid	65.02	70.89	47.9
	test accuracy (%)	softmax	55.31	56.18	56.01
		sigmoid	54.08	54.23	54.02
Output gate	content-sink mass	ungated	0.1087	0.06659	0.09815
		output gate	0.04826	0.01948	0.05598
	first-token mass	ungated	0.03589	0.03749	0.04167
		output gate	0.03793	0.03322	0.04862
	massive activation	ungated	8.026	11.68	9.215
		output gate	8.662	9.861	9.347
	perplexity	ungated	5.826	5.765	5.798
		output gate	5.807	5.803	5.761
Anchor gate	first-token mass	baseline	0.01116	0.01112	
		anchor gate	0.006846	0.006812	
	perplexity, 512 tokens	baseline	25.72	25.88	
		anchor gate	26.96	26.74	
	perplexity, 2,048 tokens	baseline	86.71	89.52	
		anchor gate	61.12	58.53	

Table 12: Every seed's value behind the comparisons of Table 3 with fewer than five seeds: sigmoid attention at depths 12 and 24 against softmax, the output gate against the ungated model, and the anchor gate against its baseline. Content-sink mass in the output-gate model rests on the value-norm test and a degenerate loading; the anchor model's first-token mass uses no stationary law.

The production-ViT census. All seven checkpoints were measured on the same 1000 ImageNet validation images, drawn with legacy `RandomState(42)` without replacement from the sorted listing and resized to 224 pixels without cropping; image identifiers index that listing. The registered predictions followed an exploratory pass over 50 images contained in the sample, so the sample is not held out. sink_cls uses strict quartile ranks over patch tokens. sink_H combines a modified z-score of the stationary mass above 3.5, a normalized inverse cluster loading above 2.5 and a value norm below the 30th percentile, on the row-renormalized,

Model	Eligible / excluded	m	S_0	S_1 (S_1^0)	S_2	$\widetilde{\psi}$ [95% CI]	S_4	Noncontr.
gpt2	144 / 0	0.536	1.000	0.968 (0.978)	0.175	−0.123 [−0.167, −0.070]	0.889	13.2%
gpt2-medium	384 / 0	0.632	0.999	0.984 (0.988)	−0.005	−0.147 [−0.181, −0.127]	0.916	10.9%
gpt2-large	720 / 0	0.579	1.000	0.981 (0.986)	−0.008	−0.175 [−0.190, −0.160]	0.916	20.1%
EleutherAI/pythia-160m	144 / 0	0.440	0.999	0.966 (0.967)	−0.356	−0.101 [−0.190, 0.027]	0.873	31.2%
EleutherAI/pythia-410m	383 / 1	0.633	1.000	0.983 (0.984)	−0.194	−0.168 [−0.203, −0.132]	0.917	24.2%
EleutherAI/pythia-1.4b	380 / 4	0.662	1.000	0.983 (0.986)	−0.134	−0.174 [−0.204, −0.140]	0.922	16.1%
facebook/opt-350m	384 / 0	0.590	0.998	0.952 (0.974)	−0.002	−0.083 [−0.151, −0.007]	0.842	3.9%
GPT-2, per input	144 / 0	0.531	0.995	0.949 (0.975)	0.042	−0.314 [−0.335, −0.290]	0.806	41.7%

Table 13: Uniform content-query weighting; conditional routing floor $r < 10^{-4}$; 95% layer-cluster intervals. First-token values are idealized as null; their value contribution is unavailable. S_4 alone uses the regularized π of an earlier census, $\epsilon = 10^{-3}$. The instrument passed its fixture checks. That earlier census is excluded: its two panels were one set of operators counted twice. Per-input validity checked.

head-averaged patch block. The stationary law comes from uniform-start power iteration (at most 500 steps, L_1 change tolerance 10^{-10}); no stationary residual was recorded, so flags in early layers are reported without interpretation. The stationary law is positive in all 96000 image–layer cells. Reversibilization is π-weighted, and because the leading right eigenvector is constant, the cluster loading is weighted by $\sqrt{\pi}$ rather than read from a trivial mode. The counterfactual arms (value zeroing and key removal) act at one layer. Source intervals resample images 400 times; the new census means use 2000 resamples. The census's predictions were registered only for ViT-B with its original ImageNet-21k pretraining, all but the check that a proxy substitutes for the ground truth, which is a robustness entry everywhere; on every other checkpoint the verdicts are robustness entries, and the channel analysis is registered in logit space and a robustness check in feature space. No record holds checkpoint hashes, revisions, or library, device or code identities; seven imported source files changed after the run, and the archived versions do not establish which of them ran. Table 14 lists all six recorded verdicts per checkpoint under their registered names. In them, with ranks over patch tokens and the registered quarter, sink_cls is the corner of tokens in the top quarter by CLS attention and the bottom quarter by composed value norm, luminous tokens are in the top quarter of both, and dark tokens are in the bottom quarter of CLS attention and the top quarter of value norm; hard_sink is the part of sink_cls that sink_H also flags, and soft_sink is the rest. Table 15 gives the change in the corner and luminous fractions between the final layer and two layers earlier and, at each rank threshold q, the Spearman correlation of the two changes across the seven checkpoints.

D.3 The package's records

Table 16 names each experiment as the text does and gives the folder of the package's `data/` directory that holds its records. The package manifest lists every file with its SHA-256 checksum, and the code in `paper_code/` builds the figures and tables from these folders.

Prediction	Verdict	Tier	Deciding quantities	Permutation null band
DeiT3-B				
lifecycle.absent_before_L5	F	robustness	worst_share=0.196	—
lifecycle.lock_then_dissolve	F	robustness	J_max=0.074	—
recruitment.dark_to_hard_sink_enrichment	F	robustness	enrichment=0.579	[0.684, 1.276]
soft_sink.previously_luminous	F	robustness	fraction=0.194, lo=0.187, hi=0.201	—
discriminator.causal_impedance_separates	F	robustness	delta=−0.126, consistent_share=0.364, layers=11	—
c3.z_proxy_substitutes_for_gt	F	robustness	layers=12, rho=0.125	—
DeiT-B				
lifecycle.absent_before_L5	S	robustness	worst_share=0.025, birth=8	—
lifecycle.lock_then_dissolve	F	robustness	J_max=0.336, dissolution=11, tenure=3	—
recruitment.dark_to_hard_sink_enrichment	F	robustness	enrichment=0.452	[0.660, 1.390]
soft_sink.previously_luminous	F	robustness	fraction=0.261, lo=0.253, hi=0.269	—
discriminator.causal_impedance_separates	I	robustness	delta=0.153, consistent_share=0.500, layers=10	—
c3.z_proxy_substitutes_for_gt	F	robustness	layers=12, rho=0.281	—
ViT-B (AugReg 1k)				
lifecycle.absent_before_L5	F	robustness	worst_share=0.437, birth=7	—
lifecycle.lock_then_dissolve	F	robustness	J_max=0.498, dissolution=8, tenure=1	—
recruitment.dark_to_hard_sink_enrichment	F	robustness	enrichment=0.279	[0.721, 1.343]
soft_sink.previously_luminous	F	robustness	fraction=0.320, lo=0.308, hi=0.333	—
discriminator.causal_impedance_separates	S	robustness	delta=0.595, consistent_share=1.000, layers=11	—
c3.z_proxy_substitutes_for_gt	S	robustness	layers=12, rho=0.558	—
ViT-B (AugReg 21k)				
lifecycle.absent_before_L5	S	robustness	worst_share=0.020, birth=8	—
lifecycle.lock_then_dissolve	F	robustness	J_max=0.801, lock=8, tenure=4, min_after_lock=0.565	—
recruitment.dark_to_hard_sink_enrichment	F	robustness	enrichment=0.271	[0.882, 1.159]
soft_sink.previously_luminous	S	robustness	fraction=0.599, lo=0.581, hi=0.614	—
discriminator.causal_impedance_separates	S	robustness	delta=0.330, consistent_share=0.583, layers=12	—
c3.z_proxy_substitutes_for_gt	F	robustness	layers=12, rho=0.288	—
ViT-B (orig. 21k)				
lifecycle.absent_before_L5	S	registered	worst_share=0.010, birth=6	—
lifecycle.lock_then_dissolve	F	registered	J_max=0.688, dissolution=10, tenure=4	—
recruitment.dark_to_hard_sink_enrichment	S	registered	enrichment=5.429	[0.912, 1.106]
soft_sink.previously_luminous	S	registered	fraction=0.465, lo=0.449, hi=0.482	—
discriminator.causal_impedance_separates	S	registered	delta=0.469, consistent_share=0.800, layers=10	—
c3.z_proxy_substitutes_for_gt	F	robustness	layers=12, rho=0.376	—
ViT-L (AugReg 21k)				
lifecycle.absent_before_L5	S	robustness	worst_share=0.015, birth=11	—
lifecycle.lock_then_dissolve	S	robustness	J_max=0.969, lock=15, dissolution=22, tenure=11, min_after_lock=0.068	—
recruitment.dark_to_hard_sink_enrichment	I	robustness	enrichment=1.955	[0.704, 1.330]
soft_sink.previously_luminous	S	robustness	fraction=0.676, lo=0.661, hi=0.691	—
discriminator.causal_impedance_separates	S	robustness	delta=0.486, consistent_share=0.667, layers=21	—
c3.z_proxy_substitutes_for_gt	F	robustness	layers=24, rho=0.381	—
ViT-S (AugReg 21k)				
lifecycle.absent_before_L5	F	robustness	worst_share=0.075, birth=9	—
lifecycle.lock_then_dissolve	F	robustness	J_max=0.942, lock=9, tenure=3, min_after_lock=0.687	—
recruitment.dark_to_hard_sink_enrichment	S	robustness	enrichment=4.750	[0.852, 1.129]
soft_sink.previously_luminous	S	robustness	fraction=0.667, lo=0.644, hi=0.690	—
discriminator.causal_impedance_separates	I	robustness	layers=0	—
c3.z_proxy_substitutes_for_gt	I	robustness	layers=0	—

Table 14: All six recorded verdicts per checkpoint, reproduced independently. Verdicts: F falsified, S supported; INCONCLUSIVE is preserved and printed as I. The recruitment band is a permutation null band, not an interval on enrichment. 1,000 images per checkpoint. Head-averaged; patch tokens only (CLS column mass is labelled separately). sink_H is the pi-outlier set; sink_cls is the CLS-readout corner. Source intervals: 400-resample cluster bootstrap over images; census means: 2,000 resamples. Lifecycle predictions are registered for ViT-B original-21k only, panel entries are robustness; channel logit space is registered and feature space is robustness. The detector differs from the one behind Table 3, and their counts are not compared.

Checkpoint / q	Δ C median	Δ luminous median	Sign class	Δ C census mean	Δ luminous census mean	Tier
DeiT3-B	0.0714	−0.0434	luminous → corner	0.0665	−0.0380	registered q=0.25
DeiT-B	0.0765	−0.0510	luminous → corner	0.0707	−0.0501	registered q=0.25
ViT-B (AugReg 1k)	−0.0306	0.0051	corner → luminous	−0.0264	0.0046	registered q=0.25
ViT-B (AugReg 21k)	0.0000	0.0357	neither	−0.0018	0.0285	registered q=0.25
ViT-B (orig. 21k)	−0.0153	−0.0102	neither	−0.0110	−0.0052	registered q=0.25
ViT-L (AugReg 21k)	−0.0510	0.0459	corner → luminous	−0.0445	0.0411	registered q=0.25
ViT-S (AugReg 21k)	−0.0357	0.0816	corner → luminous	−0.0332	0.0771	registered q=0.25
q=0.1	rho=−0.429	exact p=0.353571	n=7	—	—	robustness
q=0.2	rho=−0.750	exact p=0.066270	n=7	—	—	robustness
q=0.25	rho=−0.857	exact p=0.023810	n=7	—	—	registered
q=0.3	rho=−0.821	exact p=0.034127	n=7	—	—	robustness

Table 15: Final-layer fraction minus the fraction two layers earlier. Cross-checkpoint associations are descriptive; exact two-sided permutation p-values use midranks. No conservation claim follows. 1,000 images per checkpoint. Head-averaged; patch tokens only (CLS column mass is labelled separately). sink_H is the pi-outlier set; sink_cls is the CLS-readout corner. Source intervals: 400-resample cluster bootstrap over images; census means: 2,000 resamples. Lifecycle predictions are registered for ViT-B original-21k only, panel entries are robustness; channel logit space is registered and feature space is robustness. The detector differs from the one behind Table 3, and their counts are not compared.

Experiment	Where	Folder in `data/`
The exact construction of Figure 1	Section 1	`hero`
The numerical instance of the Jacobian bound and the Perron check	Sections 3 and 4	`abstention_parity`
Property checks of the absorbing token and the collapse law	Section 4	`RIII0`
The census of causal operators	Section 4	`RIII1`
The gauge split of trained heads	Section 5	`SX116`
The controlled pilot	Section 6, Appendix B	`pilot_logN_abstention`
Analyses of the pilot, of the retrieval and multiplier records, and of Table 3's models	Sections 6, 7 and 10	`SX40`
The robustness extension	Section 6, Appendix B	`SX41`
Sigmoid attention in a CIFAR-100 ViT	Section 7	`N19`
Four registers	Section 7	`N18`
The output gate in a character-level model	Section 7	`R7c`
The certified value-null register	Section 7	`FB10`
The anchor gate	Section 7	`AF04`
Sink mass and content energy in public language models	Section 7	`R4`
The within-head test of sink mass and contraction defect	Section 7, Appendix D	`SX42`
Sink mass as an eviction rule	Section 7	`AX31b`; an earlier run in `AX31`
Learned nulls across training lengths	Section 7	`SX112`
Positions and nulls in small byte models	Section 7	`SX113`
The production-ViT census	Section 7, Appendix D	`SX44`; its inputs in `DT-E1`, `DT-E1a`, `DT-E4`, `DT-E5a`, `DT-E14`, `DT-E15`
The harness's ViT subset	Section 8	`SX03`
Its prespecified confirmation at 896 px	Section 8	`SX102`
Attention rows at higher resolution	Section 8	`SX115`
The 125M language models	Section 8	`SX02`
Interpolation and temperature separated	Section 8	`SX114`
The test on gpt-oss	Section 9, Appendix C	`SX01`; derived slopes in `SX01_null_slopes` and `SX01_position_split`; the books audit in `SX01_books_diagnostic`
The discrete retrieval task	Section 10	`R5`, `R5b`
Continuous retrieval at length	Section 10	`R5c`, `R5d`
ViTs trained at one or several resolutions	Section 10	`V1`
Pretrained ViTs at 224 to 384 pixels	Section 10	`V1b`
The optimal logit multiplier	Section 10	`V6`
The fitted constant temperature	Section 10	`V2`
The budget in the minimal block	Section 10	`V5`
The per-head and logit-magnitude tests	Section 10	`V3`, `V4`

Table 16: Where each experiment's records are. Each folder of `data/` holds the experiment's records, with its registration where it has one.