SEPARATING OPERATOR GEOMETRY FROM LEARNED ROUTING IN A GATED, SLOT-ROUTED ASSOCIATIVE MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Associative-memory and linear-attention architectures are often diagnosed through a functional of the binding operator: key orthogonality, the spectrum of $A^\top A$, a coherence bound on the feature map. We ask what training actually changes in a gated, slot-routed associative memory, in which each stored write reaches the logit through the product of a routing factor and an operator factor. Our main instrument is *withheld learning*: one factor is frozen at its initialisation, the rest is trained, and the task's loss is measured with every slot intact. (i) *Routing.* A frozen random hard hash costs 0.1 to 1.4 accuracy points (three seeds), where a frozen diffuse router costs 37 to 55: a concentrated, balanced assignment supplies most of what routing contributes. (ii) *Writes.* Frozen write magnitudes cost 13 to 28 points (about 5 once token identity no longer marks a write's type; two loads, one seed); frozen codebook directions cost 2 to 6. (iii) *Operator geometry.* Reading the trained model agrees, once the reading is a function of the model. The customary orthogonality deviation is not: it moves by up to $255\times$ along a gauge orbit, a change of basis that leaves every logit unchanged. On a gauge-invariant contrast the operator's cross-key arrangement sits at its random-binding floor at every slot count and binding dimension we train, and no search that holds accuracy within two points lifts it past $1.03\times$. Within routing, *concentration* (how few slots a write lands on) is the lever, and *arrangement* (which writes collide) is a small, measurable channel: its mean square slot overlap sits a few per cent below permutations matched on concentration, and a fraction of one per cent below permutations matched on concentration and load. A relation joins the two readings. Under a Gaussian-crosstalk model whose variance assumption is checked, descriptors read off the forward pass (routing weights, write norms, same-key overlaps) predict retrieval accuracy to within 0.08 across 124 models with nothing fitted, so the reading is a measurement other scales can repeat rather than an extrapolation to them. The analysis stops where the memory forgets: on DeltaNet the static coefficient of a stored item does not exist, and the transition that removes it has a Lyapunov spectrum we measure. We do not offer a capacity law.

1 INTRODUCTION

A recurring move in the analysis of efficient sequence models is to explain retrieval capacity by a property of the binding operator: if keys are nearly orthogonal, writes do not interfere (Schlag et al., 2021), and the stored keys' mean pairwise cosine is read as a predictor of interference (Zahn et al., 2026); if $A^\top A$ is well conditioned, reads are clean, and the stored keys' condition number enters a lower bound on the state's effective rank (Nazari & Rusch, 2026); if the embeddings have low coherence, a Welch-type bound (Welch, 1974) caps how many pairs fit (Cabannes et al., 2024; Nichani et al., 2025; Ghriss & Chakraborty, 2026). The move is attractive because the operator is a static object, readable off a checkpoint without running the model.

This paper tests the move in a setting where it can be tested: a gated, slot-routed associative memory, one member of the linear-attention family (Katharopoulos et al., 2020; Schlag et al., 2021) and closest in form to bounded-slot attention (Peng et al., 2022; Zhang et al., 2024). The memory is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

our own construction, not a published architecture: it is the reduced form of a larger model we are developing, and it was designed independently of the routed slot memories of Afzal et al. (2026), which also route writes to a fixed set of slots. Each token writes into a distribution over M slots and reads back through a learned binding. For this memory the readout factors exactly: the coefficient with which a stored pair reaches the logit is an operator factor times a routing factor (Eq. 1). The two explanations therefore predict different things. If capacity comes from operator geometry, training should move the operator factor; if it comes from routing, intervening on routing alone should move retrieval.

Two readings that have to agree. We withhold learning, freezing one factor at its initialisation (Section 3), and read the trained model on gauge-invariant contrasts, because the customary orthogonality deviation moves $255\times$ along a gauge orbit that leaves every logit unchanged (Section 2.3). Appendix I collects the terms.

Findings at a glance.

- *Withheld learning* (Section 3): routing 0.1–1.4 points against a frozen hash, 37–55 against its diffuse initialisation; write magnitudes 13–28; codebook directions 2–6.
- *Operator* (Section 4): cross-key arrangement at its floor, at most $1.03\times$ under a search holding accuracy; what it learns is a write gate keyed to token identity.
- *Routing* (Sections 5–6): collision structure moves accuracy 10–17 points at fixed first-order statistics; the trained router arranges it in the avoiding direction by a few per cent.
- *Scaling and relation* (Sections 2.4, 7, 8): across five slot counts and binding dimensions the operator stays at its floor and the routing advantage grows with the slot count; the descriptors predict accuracy on every cell; on a delta-rule memory the per-item decomposition does not exist (Section 9).

Contributions. (1) A withheld-learning measurement, factor by factor, with every slot in place. (2) A gauge-free instrument with three matched interventions, each with its own paired null. (3) The textbook crosstalk argument (McEliece et al., 1987; Cabannes et al., 2024; Barnfield et al., 2026) instrumented on a trained model and checked with its assumptions. (4) Two failures of our own instruments, reported (Sections 6 and 9).

What we do not claim. No capacity law: the relation's skeleton is classical, and second-order overlap is the interference variance of Massey & Mittelholzer (1993) (mean square coherence, Barg et al., 2013). Ours is an intervention, not a statistic.

2 Setup: one readout, four statistics that have been confused

2.1 The memory and its readout identity

The memory writes at step s into a gate distribution $g_s \in \Delta^{M-1}$ over M slots and binds through learned codebooks $A(\cdot)$ and $B(\cdot)$. Writing $\phi(x)$ for the slot-space feature of token x, the logit for candidate word w at query step t is

$$\text{logits}_t[w] \;=\; r_t \sum_{s \le t} \omega_s \underbrace{\langle \phi(x_{s-1}),\, \phi(x_t) \rangle}_{\text{routing}} \underbrace{\big\langle A(x_t)^\top A(x_{s-1}) B(x_s),\; B(w) \big\rangle}_{\text{operator}}. \tag{1}$$

Eq. 1 is verified numerically on four trained checkpoints (worst relative error 1.07×10^{-15}). Two features differ from the idealised rank-one form: the write at step s binds the *key* at $s-1$ to the *value* at s, and the gate weighting that write is the *key's*, not the value's.

The model, completely. Three weight matrices and nothing else: A and B each map a token to an $n \times n$ matrix and ϕ maps it to a distribution over M slots, $A, B \in \mathbb{R}^{V \times n^2}$, $\phi \in \mathbb{R}^{V \times M}$, with $V = 512$, $n = 8$, $M = 16$, for 73,728 parameters — no embedding, output head, layer norm or MLP, and one block. The state is the running sum $S_t = \sum_{s \le t} \omega_s \, \phi(x_{s-1}) \otimes A(x_{s-1}) B(x_s)$, read by contracting with $\phi(x_t)$ and $A(x_t)^\top$ and projecting onto B. Here $r_t = 1$ and $\omega_s = 1$ (except

$\omega_0 = 0$): softmaxes over write and read windows of 1 in the general model from which this memory is reduced (Section 1), its other terms off and enforced off by the code.

Training. AdamW, weight decay 0.01, warmup-cosine, clipping 1.0, batch 16, 10,000 steps, 4,000 training and 500 test sequences on the official generator of Arora et al. (2024); cells $(N, T, \mathrm{lr}) = (16, 96, 10^{-2})$, $(24, 144, 10^{-2})$, $(48, 288, 3.16 \times 10^{-2})$, $(64, 384, 10^{-2})$, N stored pairs in length T; accuracy on the answer positions only.

2.2 Four statistics, named apart

One symbol has been used for four different quantities, in the literature and in our own measurements; Table 10 in Appendix I keeps them apart, together with the other terms used throughout. With s^* the step that stored the queried pair and m_s, g_s the operator and routing factors of Eq. 1, the customary orthogonality deviation $\mathrm{Orth}(A) = \|A^\top A - I\|_F / \|I\|_F$ is not gauge invariant; the operator contrast $C_{\mathrm{op}} = m_{s^*}^2 / \mathbb{E}_s[m_s^2]$ and the routing contrast $C_{\mathrm{route}} = g_{s^*}^2 / \mathbb{E}_s[g_s^2]$ are the split; and the full contrast $C_{\mathrm{full}} = c_{s^*}^2 / \mathbb{E}_s[c_s^2]$ on $c_s = m_s g_s$ is split-invariant and our primary quantity. Section 2.4 says what the split predicts.

Two specifications the numbers depend on. (a) The distractor population is the prefix's writes, grouped by distinct key, not the vocabulary: every $s \le t$, keyed by x_{s-1}, first occurrence of each key kept, of which only 37 to 40% is a stored pair. The same C_{route} reads 6.1–7.9$\times$ over stored keys and 0.36–0.78$\times$ over the vocabulary; every contrast here is over the stored keys. (b) Because ω_s in Eq. 1 is a free per-write scalar, the operator factor is also read under a cosine gauge, $\hat{m}_s = m_s / (\|A(x_{s-1})B(x_s)\| \cdot \|A(x_t)B(w^*)\|)$. Here ω_s is pinned to 1, so the raw scale is a real contribution of the operator. We report both readings; their difference is the finding of Section 4.

2.3 The customary diagnostic is not a function of the model

Eq. 1 is invariant under $A(x) \mapsto A(x)M,\ B(v) \mapsto M^{-1}B(v)$ for any $M \in GL(n)$: the operator factor contracts M against M^{-1}, so any diagnostic that is a function of the model's behaviour must be constant along this orbit. Over ten gauges — the identity, three scalars, an orthogonal, a shear, an anisotropic diagonal, three random $GL(n)$ draws — on a trained checkpoint (Table 3 and Figure 4 in Appendix B), an orbit that moves the logits by 2.31×10^{-14} moves $\mathrm{Orth}(A)$ from 0.748 to 190.635, a factor of 254.89 between the scalar gauges $M = 0.5I$ and $M = 10I$ (5.5 across the non-scalar six), while C_{op} and C_{route} do not move; transformations *outside* the group move C_{op} (to 1.803 and 1.372 from 1.624), so it is not merely blind, and nothing moves C_{route}, which does not read A or B. Canonicalising the gauge ($A \mapsto A\bar{G}^{-1/2}$ with $\bar{G}$ the mean Gram of A) makes the deviation invariant at 1.123, so the raw statistic is gauge-dependent, not the notion. A number that moves while the function is unchanged cannot be read as a property of the function — the failure Dinh et al. (2017) exhibit for sharpness under reparameterisation, here on the architecture's own group. The invariance is exact (C_{route} spread 0.0, C_{op} spread 5.1×10^{-15}).

2.4 From the three factors to accuracy

Eq. 1 makes retrieval a nearest-element decoding. At a query t with key $k = x_t$ and answer w^*, write $W_s = A(x_{s-1})B(x_s)$ for the write at position s and $g_s = \langle \phi(x_{s-1}), \phi(k) \rangle$ for its routing weight; the read state is $R_t = \sum_{s \le t} g_s W_s$, and the logit of every candidate w is its inner product with the same-key frame element $W(k, w) = A(k)B(w)$:

$$\mathrm{logits}_t[w] = \langle R_t,\, W(k,w) \rangle = g^* \langle W^*, W(k,w) \rangle + e_w, \qquad e_w = \langle N_t,\, W(k,w) \rangle, \tag{2}$$

with $W^* = W(k, w^*)$ the true write, g^* its routing weight and $N_t = R_t - g^* W^*$ the rest of the state. Retrieval is correct iff $g^* \|W^*\|^2 (1 - \rho_w) + e_{w^*} - e_w > 0$ for every $w \neq w^*$, where $\rho_w = \langle W^*, W(k,w) \rangle / \|W^*\|^2$ is the same-key overlap profile of the frame at k, a property of the codebooks alone. Nothing so far is approximate. The classical step (McEliece et al., 1987; Massey & Mittelholzer, 1993) treats the crosstalk e_w as Gaussian with the independent-interference variance

$$\mathrm{Var}\, e_w = P_N \, \|W(k,w)\|^2 \, c_d, \qquad P_N = \sum_{s \neq s^*} g_s^2 \|W_s\|^2, \tag{3}$$

c_d the mean square cosine between a frame element at k and a distractor write, every prefix write counted. Two assumptions are made: (i) e_w is Gaussian over the draw and, given the common true-logit term e_{w^*}, independent across competitors and of e_{w^*}; (ii) cross terms between distractor writes are dropped. (ii) is checked directly in Section 8; (i) only through the prediction it produces (Appendix F, with and without its independence clause). Under (i)–(ii) the probability of a correct read is a one-dimensional integral, and its union bound is a bound under any dependence between competitors:

$$P_t = \mathbb{E}_{e^*}\Big[\prod_{w\neq w^*}\Phi\Big(\tfrac{g^*\|W^*\|^2(1-\rho_w)+e^*}{\sigma_w}\Big)\Big], \qquad P_t \;\geq\; 1-\sum_{w\neq w^*} Q\Big(\tfrac{g^*\|W^*\|^2(1-\rho_w)}{\sqrt{\sigma_*^2+\sigma_w^2}}\Big). \tag{4}$$

The signal-to-noise ratio behind it factors into the three objects this paper measures,

$$\mathrm{SNR}_t = \underbrace{\frac{g^{*2}}{\sum_{s\neq s^*} g_s^2}}_{\text{routing}}\;\underbrace{\frac{\|W^*\|^2}{\overline{\|W_s\|^2}_{g^2}}}_{\text{write gate}}\;\underbrace{\frac{(1-\bar\rho)^2}{c_d\,(1+\bar\nu^2)}}_{\text{frame}}, \tag{5}$$

with $\overline{\|W_s\|^2}_{g^2}$ the g_s^2-weighted mean distractor norm, $\bar\rho$ the mean same-key overlap and $\bar\nu^2$ the mean competitor norm relative to $\|W^*\|^2$: the routing factor is C_{route} per interferer, the write-gate factor the magnitude part of C_{op}, the frame factor the same-key and cross-key coherence of Section 4. Every input is a forward-pass quantity; Section 8 evaluates the relation, with nothing fitted, on every model of Sections 3–7 and on each coordinate with nothing trained.

3 What training changes, measured by withholding it

The statistics that follow read a trained model after the fact. Before reading anything, we ask the question directly: freeze one factor at its initialisation, train the rest under the same data, budget and seed, and measure what the task loses; every slot stays in place in every arm, so a frozen router is a *random* routing over sixteen slots, not a collapse to one. Four numbers come out of Figure 1a (per seed in Appendix F). Learned routing is worth 0.1 to 1.4 accuracy points against a frozen random hard hash, one slot per symbol (Roller et al., 2021), on three seeds, resolved above the trained model's seed spread only at $N = 48$; against its diffuse initialisation, every slot intact, routing is worth 37 to 55 (Appendix F): a concentrated, balanced assignment supplies most of what routing contributes, and what learning adds beyond a hash is of the size of the arrangement channel of Section 6. Learned write magnitudes are worth 13 to 28 (about 5 without the identity mark, Section 10): freezing the codebooks entirely costs 15 to 33, and letting one scale per row train recovers all but 2 to 6. Learned codebook directions — everything a per-row scale cannot express — are worth those last 2 to 6. The router arm withholds the routing factor of Eq. 5, the frozen-codebook arm the write-gate factor (and, through $\bar\nu^2$, the frame factor); freezing directions acts through the same-key overlap profile of Eq. 4. A learning-rate sweep of every arm moves no cost by more than 1.6 points (Appendix F).

A condition number is the second diagnostic that runs the wrong way. Figure 1b reads the state a query decodes from with each component switched on or off. Geometry reads the same on all four arms; what moves the conditioning is the gate (4 to $48\times$) and the routing (43 to $340\times$). Unlike Orth it is a function of the model, and still not a quality: the model as it reads is nearly three orders worse conditioned than its initialisation, at the highest accuracy in the paper, DeltaNet inverts the same way, and the routing-over-gate ordering is readable at three depths of a released 1.3-billion-parameter memory trained on text (Appendix G).

4 The operator: training changes its scale, not its cross-key arrangement

Training *does* change the operator: the raw contrast rises 2.1 to $2.6\times$ (Table 1). It does not change the cross-key arrangement the contrast reads: once the per-write scale is removed the contrast sits at 0.997 to 1.002 times the floor, with a bootstrap 95% interval on the median ratio inside $[0.988,\ 1.010]$ over 4096 paired draws (cross-seed spread about 7%, Appendix F). Those draws resolve a true ratio of 1.014 to 1.021, inside the 4.1 to 6.6% of room the frame bound leaves at the

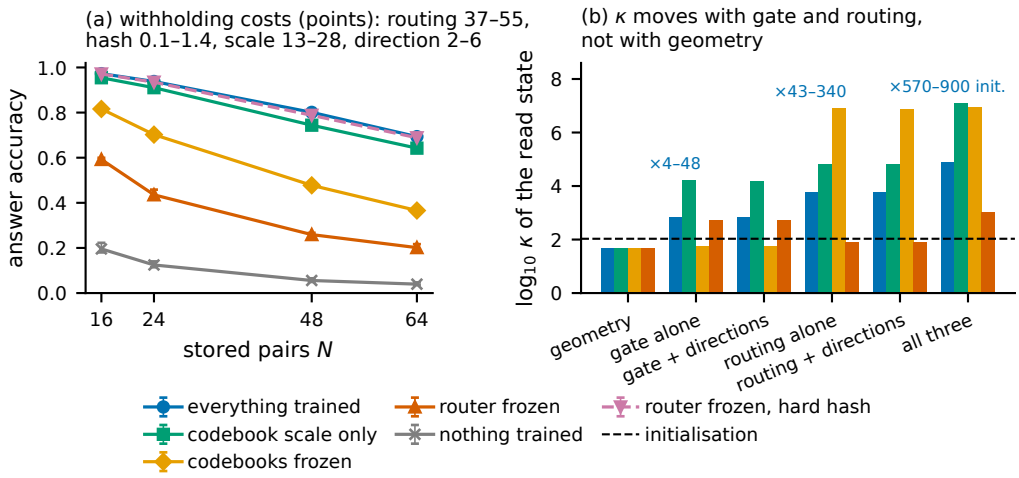

Figure 1: Withholding one factor from training costs most for routing against its diffuse initialisation, then for write scale, least for codebook direction; a frozen random hard hash costs 0.1 to 1.4, below the direction cost; the condition number of the read state moves with the gate and the routing, not with the geometry. (a) Answer accuracy with one factor withheld, three seeds (bars: seed range; per seed in Appendix F), and with the router frozen as a random hard hash, three seeds (purple, dashed; bars: seed range; it lies under the everything-trained curve, the gap being its cost). (b) Condition number of the read state's Gram with each component switched on alone, on random or trained directions, and all together; medians over queries, mean of $\log_{10}\kappa$ over loads (per load in Appendix F); dashed: initialisation; annotated: the everything-trained arm's ratios. The frozen-codebook arms route hardest and read highest.

Table 1: Training raises the operator contrast 2.1 to 2.6× through the per-write scale alone: in the cosine gauge the trained/floor ratio is 0.997 to 1.002 (bold), while the routing contrast rises 6 to 8×. Trained over a paired random-binding floor on the official generator of Arora et al. (2024), medians over 400 draws per cell (cosine column: 4096).

pairs N	accuracy	C_{op} raw	C_{op} cosine	C_{route}	C_{full}
16	0.9717	2.58	**0.997**	6.06	19.69
24	0.9378	2.50	**1.002**	6.68	14.28
48	0.8000	2.28	**1.000**	7.62	13.63
64	0.6909	2.06	**1.002**	7.93	13.15

trained same-key overlap (Appendix D), so this is a bounded non-detection and not the design failing to look. The contrast reads the target against the distractors along $B(w^*)$ and is blind to same-key separation, which training does move: the mean same-key overlap falls by 35% (Appendix D).

What the raw rise is, then. Not a larger queried value (the queried pair is written no harder than any other true pair, within 0.7%): what training learns is which writes are *real*. A write at s binds the key at $s-1$ to the value at s, so the sequence produces true pairs, *cross* writes binding a value to the next key, and query-side writes, and 64% of the distractor population is the latter two, artefacts of the shift. Training separates them by norm: true pairs are written 1.9 to 2.8 times harder than at initialisation, cross writes 2.1 to 4.4 times softer, query-side writes 1.6 to 2.1 times softer (Appendix F); the type is read from token identity alone (Section 10). Removing that gate — renormalising every write to unit Frobenius norm — costs 6.5 to 20.1 accuracy points, against 66.2 to 85.9 for replacing every routing weight by its mean; the two ablations destroy different amounts, so their ratio is not the factors' intrinsic shares, which is why Section 3 withholds instead. Meanwhile C_{route} rises 6 to 8-fold and C_{full} 13 to 20-fold: the gain is in the routing factor (on true stored pairs alone, 4.0 to 4.7-fold).

Was there anything to take? A value at the floor says nothing until the reachable range is known, so we optimise A and B directly against the cosine-gauge contrast — no task, no labels — and ask where it saturates (Appendix F). Optimised and read on the same sequences it climbs to $1.76\times$ the floor ($65{,}536$ parameters fitted to a few hundred rows; on disjoint sequences 0.99 to $1.01\times$). Resampling the objective at every step removes the set there was to fit, and then the contrast *does* leave the floor, to 1.38 to $2.03\times$ on a disjoint reading set — at a cost of 21.7 to 26.6 accuracy points, part of it the step size (Section 10); under an accuracy constraint no setting in our grid buys more than $1.03\times$ while staying within two accuracy points, and at the rates where the search does find contrast nothing that satisfies the constraint lifts it (Section 10). A frame potential (Benedetto & Fickus, 2003) on unit-norm writes in $d = n^2 = 64$ dimensions (Appendix D) bounds the room among stored pairs at 5.4 to 11.2% with the same-key overlap held, and the search does not hold it: at its own optimum the same-key mean square cosine is 1.78 to 1.85 times the trained binding's and above the random binding's on all four checkpoints, and on stored pairs alone its contrast is 0.81 to 0.90 of the floor: the gain sits in the artefact writes (Appendix D). That a random, non-orthogonal binding is already near optimal is what Nichani et al. (2025) prove for linear associative memory, and Cabannes et al. (2024) and Vural et al. (2026) derive how such memories scale with non-orthogonal embeddings; ours is the measurement beside them — the operator's cross-key arrangement is not a lever here.

Dong et al. (2025) (frozen query and key projections) and Bietti et al. (2023) give the model our finding instantiates, coefficients growing on a fixed near-orthogonal geometry; Tay et al. (2021) points the other way (3.4 BLEU on En-De for fixing the Random Synthesizer's alignment).

5 Routing I: the first moment does not determine retrieval

Under uniform random addressing the second moment of the overlap follows from the first, which is why classical analyses of distributed memory fix the mean overlap; a learned soft router is not under that constraint. We construct two gate families over $M = 16$ slots with *identical* first-order and per-item statistics — mean pairwise overlap, effective slot count $\exp H$, participation ratio $1/\sum_i g_i^2$, number of distinct supports — that differ in the second-order collision structure (supports on a grid against cyclic shifts). The effect is 10.4 to 17.2 accuracy points (Figure 5 in Appendix C), against a three-seed reproduction band of 0.0049 and an evaluation-draw spread of 0.017 (Reproducibility Statement).

Second moment, or the worst overlap riding along with it? The two pure families also differ in the maximum overlap ($2/16$ against $3/16$), so worst-case coherence (Mixon et al., 2011) moves with the second moment. A five-rung ladder separates them (Table 4 in Appendix C): mixing the families at fixed proportions holds support size fixed while varying the collision structure; the mean overlap is matched exactly at two pairs of rungs, 0 against 16 and 4 against 12, both carrying a gap of $+0.10$ to $+0.17$, and rungs 4 and 12 also share the maximum while differing in the second moment by $1.39\times$. Ordered by second moment, accuracy tracks it (rank correlation ≤ -0.90 on four of four checkpoints); the mixture that collides most (twelve shifts) is the least accurate and accuracy returns at sixteen by $+0.030$ to $+0.067$. The hash router of Section 3 is no rung: one-hot, its $\mathbb{E}[b^2]$ is 0.061, five to ten times the trained routers', at almost their accuracy; the ladder compares gates of one support size, and a change of concentration is read by Eq. 5's routing factor, 16 for the hash against 22 to 29 trained and 3 to 4 at the floor (Appendix F).

6 Routing II: the router does arrange collisions, and our first statistic could not see it

Section 5 shows the lever exists; this section asks whether training pulls it. Our first answer was no, and the way it was wrong is the useful part. The freedom is large: at a fixed effective slot count (3.596) and participation ratio (3.333), C_{route} spans $19.27\times$.

The right null, and the right statistic on it. We pair each trained gate table against a *per-symbol slot permutation* of itself, which preserves every symbol's own concentration exactly and destroys any arrangement *between* symbols. On the mean square coherence against 200 such permutations

per checkpoint, weighted by how often the generator presents each symbol as a stored key, the trained router sits 4.4% to 5.9% below every one ($z = -2.3$ to -3.0), against a closed form matching the 200-draw mean to 0.5%. Two things qualify it. The router balances load without being asked — within 0.04 to 0.46% of the uniform minimum $1/M$ with no auxiliary loss, against 2 to 5% above it for its own permutations — and a permutation does not hold that load, so the gap is an upper bound. Removing the first-order load term leaves 4.0 to 4.9% (rank 0 of 200); a null annealed to hold concentration *and* load to 10^{-4} leaves 0.2 to 0.6% (rank 0 of 50, $z = -2.3$ to -4.0), conservative because it destroys some arrangement while matching (Table 6, Appendix F).

Our first pass read this comparison as a null: not the statistic but its *estimate*, sampled over symbol subsets with a per-draw spread some thirty times the effect, where the exact computation costs one 512×512 product.

The size of the channel. Small next to concentration, and not nothing: flattening the gate by a temperature costs 0.774 mean accuracy, concentrated in the two doublings past the trained operating point; destroying arrangement at fixed concentration costs 0.005 to 0.031 mean accuracy over 200 permutations, and 0.001 to 0.011 once the load is matched as well (Appendix F); sharpening buys nothing. Beside Ahrac et al. (2026) (routers and experts learning a coupled geometry) ours is the narrower claim, consistent with Roller et al. (2021), that collision structure moves by a few per cent; on the mixture of memories of Du et al. (2026), run here as a diagnostic with its balancing devices off, the router collapses onto one memory and leaves only load to read (Appendix G).

7 Scaling in slots and in binding dimension

Everything above is at $M = 16$, $n = 8$; the same four loads were retrained at $M \in \{8, 32\}$, $n = 8$ and at $n \in \{4, 16\}$, $M = 16$, and every statistic re-read on them (Table 8 in Appendix F).

Three things hold across the sweep. The operator contrast sits at its floor at every coordinate, so the null of Section 4 is not a property of $M = 16$, $n = 8$. The routing advantage grows with the slot count, $4.6\times$ at eight slots to $14\times$ at thirty-two, and so does the arrangement effect. And the binding dimension moves accuracy far more than any operator statistic — $n = 4$ loses two thirds of the task at $N = 48$ with C_{op} at 1.0 — as Eq. 5 says it should: $d = n^2$ enters the frame factor through $c_d \approx 1/d$, not through anything training moves.

8 The relation, checked: the descriptors predict every accuracy above with nothing fitted

Eq. 4, evaluated from the descriptors alone, is set beside measured accuracy on every cell this paper trains or initialises: the arms, gates and coordinates of Sections 3–7, and every coordinate read with nothing trained, at each seed's initialisation (Figure 2; per group in Table 7). Assumption (ii) holds: the measured crosstalk variance is 0.94 to 1.00 times Eq. 3 on every trained cell (0.87 to 0.93 on the $n = 4$ initialisation cells); that checks the variance, not the Gaussian shape, which (i) supplies. Nothing is fitted: the descriptors are read off the forward pass and Eq. 4 evaluated, no parameter tuned on the accuracies it meets. The prediction lands within 0.08 of measured accuracy on all 112 cells, and on the twelve hash-router cells, where $g_s \in \{0, 1\}$ (by at most 0.059, Appendix F) (mean -0.03; 0.054 with the covariance of Appendix F), below the measurement on 101 of them and above it by at most 0.011 on the other eleven (Table 7); it preserves the order of the arms, of the coordinates and of the rungs (rank correlation 0.999), and the bound exceeds the measurement on two cells, by at most 0.006. The under-prediction peaks at mid accuracy; the prediction holds with the argmax restricted to 64 candidates (Figure 2c). Its inputs are the per-query descriptors behind Eq. 5 (routing weights, write norms, ρ_w), not the three scalars, whose product misses by up to 0.34 (Appendix F). It is not a capacity law, but it makes those descriptors sufficient for retrieval under (i)–(ii).

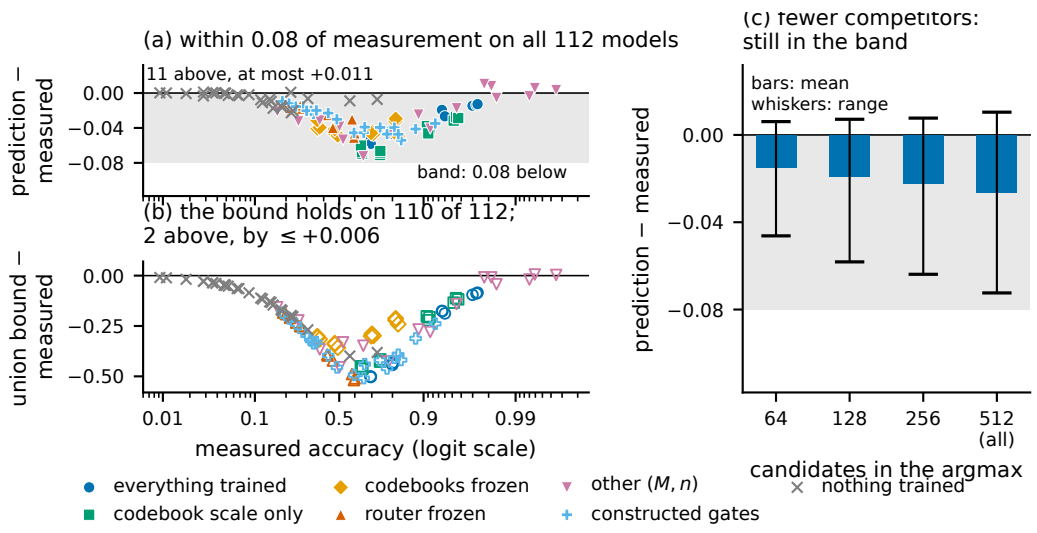

Figure 2: The forward-pass descriptors predict accuracy to within 0.08 on every model this paper trains or initialises, and the union bound holds on all but two. (a) Prediction of Eq. 4 minus measured accuracy on 112 models, grouped as in the legend; band: 0.08 below the measurement; seed replicates overlap. (b) The union bound minus measured accuracy, same models. (c) The argmax restricted to 64, 128 and 256 random candidates on the same checkpoints: mean and range of the error, on the scale and band of (a).

9 Where this analysis stops: memories that forget

We pre-registered a scope test, expecting a positive control, on DeltaNet (Yang et al., 2024b), where the write is a retrieve-and-subtract update (Schlag et al., 2021) $\tilde{v}_s = \beta_s(v_s - S_{s-1}^{\top} k_s)$ (one layer, two heads, vocabulary 8192; Appendix G). It failed: the margin is negative on 84.0% to 94.5% of queries at 8 to 64 pairs, on checkpoints that answer correctly 99.93% down to 52.47% of the time (Appendix G). The target window holds the largest contributor on every query and still loses to some distractor group, because the quantity written at step t, $u_t = \beta_t\big(v_t - \sum_{i<t} u_i(k_i^{\top} k_t)\big)$, is a residual that depends on the whole prefix. We stopped at the three definitions the preregistration allowed: adjusting a definition until the margin turns positive is what it exists to prevent.

The scope condition has a parameter. Eq. 1 is static because the memory's transition between writes is the identity; with $P_t \neq I$ each stored item still has a coefficient, but one depending on every key after it, sized by the Lyapunov spectrum of the transition product (Oseledets, 1968) (Figure 3b). Ours is 0; GLA, Mamba-2 and RetNet, at their published rate, forget at 0.03 to 1.2 per token; DeltaNet, driven by its own trained keys, holds its slowest channel near identity within a sequence ($\lambda_1 = -0.0002$ to -0.0007, six to thirty-two times slower than isotropic keys give) and forgets fastest at 0.06 to 0.41, in four directions carrying 4 to 14 times their share of key energy, so erasure is confined to a low-dimensional channel (Appendix G). A second exact decomposition, run once, puts the erasure into that coefficient and makes the read attributable at $N = 8$ (margin positive on 67% of queries, against 16%) and not at $N \geq 32$ (25 to 34%); load and head dimension change together there ($d_h = 64$ at $N = 8$, 32 at $N \geq 32$), so the boundary is not placed. What survives is the decoding (Figure 3a): split against the convolution window and anchored at the true write, Eq. 4 predicts DeltaNet's accuracy within 0.10 from below at every load and within 0.06 on every withheld arm — a prediction, not a result, the variance ratio of (ii) reading 0.3 to 0.6 there. Preregistered withholding gives the first element of the order of Section 3 against a diffuse initialisation (frozen addressing at 0.04 or below, a frozen erasure gate at most 2.3 points), and so does the fingerprint under all three value metrics: addressing moves the conditioning 4× to 33× where the statistic resolves, the erasure gate at most 4.1× (Appendix G; on the delta gate Hatamizadeh et al. (2026), on attribution in recurrent models Rezaei Jafari et al. (2024)).

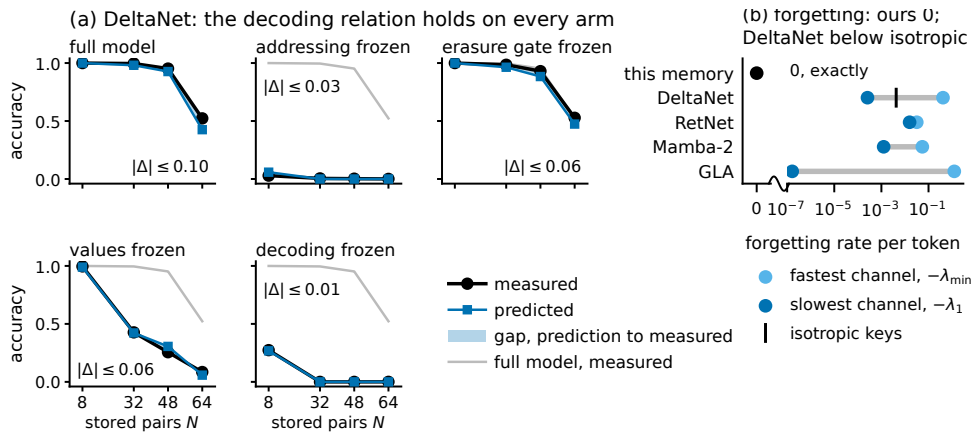

Figure 3: On DeltaNet the decoding relation predicts accuracy within 0.10 on the full model and within 0.06 on every withheld arm; DeltaNet forgets where this memory does not. (a) Accuracy against the load, one panel per arm: measured (black), predicted by Eq. 4 with the read split against the convolution window (blue), the gap shaded and its maximum $|\Delta|$ printed, the full model in grey. (b) Slowest and fastest forgetting rates per token ($N = 8$): GLA, Mamba-2, RetNet from their trained decays (RetNet below the certification gate); DeltaNet from its transition's Lyapunov spectrum, with the isotropic-key rate for its slowest channel; this memory's rate is exactly zero (Appendix G).

10 Limitations

Other scales. Every routed-memory model has $V = 512$, $d \leq 256$, $M \leq 32$ and one load ray, $N/T = 1/6$ (DeltaNet: $V = 8192$, N/T $1/16$ to $1/8$). The identities, frame bound and null do not depend on that; the relation is validated only there, V only by restricting the candidates (Figure 2c), and assumption (i)'s tail only at z of 2 to 4.

Other data. Only the accuracy check needs ground truth; the text case is not run. Keys and values come from disjoint halves of the vocabulary, so token identity marks a write's type and the gate of Section 4 reads it; from one pool the gate is gone (norm ratio 0.98 to 1.01 against 4.1 to 12.2), accuracy falls to 0.24 to 0.78, two loads do not converge, the magnitude cost falls to about 5 points and the direction cost rises to 4 to 5 (Appendix F).

The constructed gates are hard-uniform: the ladder shows the lever, not that the gap appears among gates a trained router visits; no pair of rungs isolates $\mathbb{E}[b^2]$.

The reachable range is an optimisation result. Descending the task loss alone at the same rates also costs accuracy: at $N = 16$ more than the search, at $N = 48$ and 64 2 to 8 of its 11 to 19 points (Appendix F); one optimiser.

11 Conclusion

Withholding learning shows that routing (37 to 55 points over its initialisation, 0.1 to 1.4 over a random hash), write magnitudes (13 to 28) and codebook directions (2 to 6) carry what training adds, while the operator's cross-key arrangement stays at its floor; the descriptors predict accuracy within 0.08 across 124 models.

Implications. Within this regime (synthetic recall, $V = 512$): a fixed, balanced random hash recovers the learned router to within 1.4 points, so routing needs concentration and balance rather than a learned assignment; the operator's arrangement sits at its floor and raw $\mathrm{Orth}(A)$ moves $255\times$ with the function unchanged, so diagnostics belong on gauge-invariant contrasts.

Reproducibility Statement

Seeds. The four cells and the withheld arms were retrained at two further seeds: accuracy spreads by at most 0.0049 across the three and 0.036 on the arms, against 0.017 from evaluation draws alone, and every contrast lands inside that range or beside it. The ceiling, sweep and ladder are seed 17 only; a seed range checks reproduction, not significance (Appendix F).

Equation 1 is verified numerically in the supplementary artifact (worst relative error 1.07×10^{-15}). Every routed-memory measurement uses the official generator of Arora et al. (2024), pinned by commit hash in the artifact's provenance record; the harness refuses to run if that package is not importable, because an earlier version of this work silently fell back to a local port. The DeltaNet, GLA, Mamba-2, RetNet and MoM cells of Section 9 were trained and read on the artifact's numerical port of that generator (Appendix G). The preregistration — readout direction, distractor grouping and the paired floor construction — is reproduced in Appendix E and was committed before the measurements it governs; the stopping rule for the positive control was adopted after its first attribution and is dated there. Every number in every table is emitted by a script into a JSON file of the same name (the full model's training-time accuracies of Table 1 and the seed table come from the training log; the same checkpoints replay at 0.9699, 0.9363, 0.7998, 0.6931 on the relation's draw), and the figures are generated from those JSON files; the scripts and JSON files ship with the artifact.

The four seed-17 checkpoints at $(M, n) = (16, 8)$ and the four DeltaNet checkpoints ship with the artifact; the seed-29 and 41 cells, the withheld arms, the sweep and the baselines are retrained by the shipped scripts. Tensors are bit-identical to the ones used here, verified by hash, and a shipped script replays the relation's test draw on the packaged checkpoints and reproduces the four recorded accuracies exactly.

Ethics Statement

This work trains small synthetic-recall models and analyses their internals. It involves no human subjects, no personal data, and no released model weights capable of harm. The architectural finding is diagnostic rather than capability-increasing.

The Use of Large Language Models

This paper was produced in a sustained collaboration with a large language model (Anthropic's Claude, used through an agentic coding interface). Its role was substantial rather than editorial, and so was the authors' role in directing, checking and correcting it. Table 2 states both by research subtask; the paragraphs after it say how the authors directed the model, what the model proposed, and the errors that were found and corrected.

Table 2: Division of work by research subtask.

Subtask	The model	The authors
Research question	Did not choose it.	Chose it, and redirected it as results came in; set the claims the paper makes.
Protocol	Followed it.	Fixed the sequence every experiment went through: literature survey, design, smoke test, pilot, main run, each step approved before the next.
Design	Proposed the instruments, controls and arms listed below.	Accepted, rejected or redirected each proposal; obtained the outside review that asked for four of the arms.
Code	Wrote essentially all of it: every measurement script, the figure code and the anonymisation gate applied to the supplementary artifact, generally from a one-sentence specification of what needed to be established.	Specified what each measurement had to establish.
Running experiments	Executed the scripts and collected the JSON artifacts.	Authorised every run; no experiment ran without explicit authorisation.
Derivation	Derived the relation of Section 2.4 and wrote its evaluation.	Asked for it, as a way of making the paper's readings testable at other scales; decided how the paper's theory is stated and presented.
Tables and figures	Drafted them from the JSON artifacts.	Reviewed every result against its artifact, and every figure for what it shows and how it reads, before inclusion.
Writing	Drafted the prose, and edited it in revision.	Set the paper's structure, corrected the prose and its context, and decided what enters the paper; the model had no authority to accept a claim.
Checking	Ran adversarial reviews of its own derivation, code and prior-art coverage with several parallel instances of itself; one found the double-counted variance term and another the test-set discrepancy (Appendix H).	Found errors in the model's work and overturned judgments it had made; had every review finding verified against the artifacts before any edit.

How the authors directed the model. The model worked under standing rules the authors set and enforced. Every experiment followed the authors' protocol: a survey of the literature and of existing code, then a design, then a smoke test run on the real analysis path, then a pilot, then the main run, with no step begun before the previous one was approved. No run started without their prior approval and an estimate of its duration; every criterion fixed in a dated declaration before the data it governs; every feedback item, including the model's own, verified against the artifacts before it changed the paper; every reported number traceable to an artifact; and every error disclosed rather than repaired quietly. The authors obtained reviews from outside readers and from panels of model instances, put each finding to the model as a task, and decided which findings the paper would act on. They challenged conclusions the model had reached, found errors it had not, and overruled judgments it had made; they set the paper's structure, corrected its prose and context, and reviewed its figures. They decided which claims the paper makes and how its theory is presented. The questions the model answered, and the standard its answers were held to, were the authors'.

What the model proposed. Each item below was proposed by the model and accepted by the authors. *Design:* the gauge-orbit experiment; the per-symbol permutation as the concentration-matched null; keeping the failed positive control in the paper rather than deleting it. *In response to critical reading:* the held-out and then fresh-draw ceilings; the split of writes by type; the write-gate ablation; the frequency-weighted arrangement population; the worst-overlap control on the ladder. *In later revisions:* the withheld-learning arms; the accuracy-constrained search and the resampling ceiling read on stored pairs; the load-matched and centred arrangement nulls; the slot-count and binding-dimension sweep; the stable-rank and tolerance reading beside every condi-

tion number; the fingerprint read at three depths of a released 1.3-billion-parameter checkpoint; the Lyapunov and stationary-state analysis of DeltaNet with its three hypothesis tests; the forgetting axis across four other memories with its learning-rate sweep, trained through the official `flash-linear-attention` library (Yang & Zhang, 2024) and MoM packages (the latter with a one-line signature patch, disclosed in the artifact). Four arms were asked for by a review from outside this collaboration, not one of the model's own panels, which the authors obtained and put to the model: the learning-rate sweep of the withheld arms, the frozen hard-hash router, the shared-vocabulary variant with its two codebook arms, and the task-loss-only control.

Errors the model made and corrected. Appendix H records every correction and is part of this disclosure. Errors in the model's work were found by its own checks and by the authors' reading, and corrected, some after they had reached a draft. Among them: a routing contrast measured against the full vocabulary instead of the sequence's stored keys, which inverted the sign of the effect; a power claim of "18 standard errors", withdrawn because it divided two quantities on different scales; a single-draw estimate of a heavy-tailed ratio, unrepresentative by an order of magnitude and replaced by a median over many draws; a DeltaNet extraction that omitted the model's short convolution, caught only because the resulting margin was negative on a 99.96%-accurate model; and a harness that silently used a local port of the generator instead of the official package, after which every number was re-measured. An earlier version also reported a power curve for what was then the central negative result; the model later showed that result to be an artefact of its own estimator, and both the curve and the claim are withdrawn. On one occasion the model incorrectly voided a correct result of its own, and retracted the voiding when a later check proved the original right.

Responsibility. The authors take full responsibility for the contents of this submission, including for any error the process above failed to catch.

REFERENCES

Arshia Afzal, Aviv Bick, Eric P. Xing, Volkan Cevher, and Albert Gu. Raven: High-recall sequence modeling with sparse memory routing. *arXiv preprint arXiv:2607.25357*, 2026.

Sagi Ahrac, Noya Hochwald, and Mor Geva. Routers learn the geometry of their experts: Geometric coupling in sparse mixture-of-experts. *arXiv preprint arXiv:2605.12476*, 2026.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*, 2024.

Alexander Barg, Arya Mazumdar, and Rongrong Wang. Random subdictionaries and coherence conditions for sparse signal recovery. *arXiv preprint arXiv:1303.1847*, 2013.

Nicholas Barnfield, Juno Kim, Eshaan Nichani, Jason D. Lee, and Yue M. Lu. Sharp capacity thresholds in linear associative memory: From top-1 retrieval to tail-average learning. *arXiv preprint arXiv:2605.05189*, 2026.

John J. Benedetto and Matthew Fickus. Finite normalized tight frames. *Advances in Computational Mathematics*, 18(2–4):357–385, 2003.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, 2023.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *International Conference on Learning Representations*, 2024. arXiv:2310.02984.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, pp. 10041–10071, 2024.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028, 2017.

Yihe Dong, Lorenzo Noci, Mikhail Khodak, and Mufan Li. Is random attention sufficient for sequence modeling? disentangling trainable components in the transformer. *arXiv preprint arXiv:2506.01115*, 2025.

Jusen Du, Weigao Sun, Disen Lan, Jiaxi Hu, Tao Zhang, and Yu Cheng. MoM: Linear sequence modeling with mixture-of-memories. In *International Conference on Learning Representations*, 2026. arXiv:2502.13685.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 2022.

Ayoub Ghriss and Sourav Chakraborty. Kernelized linear attention: Breaking the capacity wall with symmetric cones. *arXiv preprint arXiv:2607.17419*, 2026.

Ali Hatamizadeh, Yejin Choi, and Jan Kautz. Gated deltanet-2: Decoupling erase and write in linear attention. *arXiv preprint arXiv:2605.22791*, 2026.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, 2020.

Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Large memory layers with product keys. In *Advances in Neural Information Processing Systems*, 2019.

James L. Massey and Thomas Mittelholzer. Welch's bound and sequence sets for code-division multiple-access systems. In *Sequences II: Methods in Communication, Security and Computer Science*, pp. 63–78. Springer, New York, 1993.

Robert J. McEliece, Edward C. Posner, Eugene R. Rodemich, and Santosh S. Venkatesh. The capacity of the Hopfield associative memory. *IEEE Transactions on Information Theory*, 33(4): 461–482, 1987. doi: 10.1109/TIT.1987.1057328.

Dustin G. Mixon, Waheed U. Bajwa, and Robert Calderbank. Frame coherence and sparse signal processing. *arXiv preprint arXiv:1105.4279*, 2011.

Philipp Nazari and T. Konstantin Rusch. The key to state reduction in linear attention: A rank-based perspective. *arXiv preprint arXiv:2602.04852*, 2026.

Eshaan Nichani, Jason D. Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In *International Conference on Learning Representations*, 2025. arXiv:2412.06538.

Valery Iustinovich Oseledets. A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems. *Transactions of the Moscow Mathematical Society*, 19:197–231, 1968.

Hao Peng, Jungo Kasai, Nikolaos Pappas, Dani Yogatama, Zhaofeng Wu, Lingpeng Kong, Roy Schwartz, and Noah A. Smith. ABC: Attention with bounded-memory control. In *Annual Meeting of the Association for Computational Linguistics*, 2022. arXiv:2110.02488.

Farnoush Rezaei Jafari, Grégoire Montavon, Klaus-Robert Müller, and Oliver Eberle. MambaLRP: Explaining selective state space sequence models. *arXiv preprint arXiv:2406.07592*, 2024.

Stephen Roller, Sainbayar Sukhbaatar, Arthur Szlam, and Jason Weston. Hash layers for large sparse models. *Advances in Neural Information Processing Systems*, 34:17555–17566, 2021.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention for transformer models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 10183–10192, 2021.

Nuri Mert Vural, Alberto Bietti, Mahdi Soltanolkotabi, and Denny Wu. Learning to recall with transformers beyond orthogonal embeddings. In *International Conference on Learning Representations*, 2026.

Lloyd R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Transactions on Information Theory*, 20(3):397–399, 1974.

Songlin Yang and Yu Zhang. FLA: A Triton-based library for hardware-efficient implementations of linear attention mechanism. https://github.com/fla-org/flash-linear-attention, 2024.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *International Conference on Machine Learning*, pp. 56501–56523, 2024a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *Advances in Neural Information Processing Systems*, 2024b. arXiv:2406.06484.

Oliver Zahn, Matt Beton, and Simran Chana. Attention is not retention: The orthogonality constraint in infinite-context architectures. *arXiv preprint arXiv:2601.15313*, 2026.

Yu Zhang, Songlin Yang, Ruijie Zhu, Yue Zhang, Leyang Cui, Yiqiao Wang, Bolun Wang, Freda Shi, Bailin Wang, Wei Bi, Peng Zhou, and Guohong Fu. Gated slot attention for efficient linear-time sequence modeling. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2409.07146.

A THE READOUT IDENTITY AND ITS VERIFICATION

Let $\phi(x) \in \Delta^{M-1}$ be the gate distribution written by token x, and let $A(\cdot), B(\cdot)$ be the key and value codebooks. Accumulating writes into the slot state and contracting at query time gives Eq. 1 directly. Two features of that expression are worth isolating because both differ from the rank-one form usually assumed when this family is analysed, and an appendix that assumed the usual form would produce a different — and wrong — operator factor.

1. **The write at step s binds the key at $s-1$ to the value at s.** The operator factor is therefore $\langle A(x_t)^\top A(x_{s-1})B(x_s), B(w)\rangle$ and not $\langle A(x_t)^\top A(x_s)B(x_s), B(w)\rangle$.
2. **The gate weighting that write is the key's gate.** The routing factor is $\langle \phi(x_{s-1}), \phi(x_t)\rangle$, indexed at $s-1$.

Verification. We evaluate the right-hand side of Eq. 1 term by term and compare against the logits produced by the running model, on four trained checkpoints. Worst relative error over all four: 1.0683×10^{-15}, on the official generator. The routing factor cancels exactly out of any ratio taken at fixed t, and the operator factor is w-dependent — both are consequences of the identity rather than assumptions made before it.

B THE GAUGE ORBIT

The group is $A(x) \mapsto A(x)M$, $B(v) \mapsto M^{-1}B(v)$. The ten gauges are: identity; scalar $c \in \{0.5, 2, 10\}$; a random orthogonal Q; the shear $I + 3e_0e_1^\top$; an anisotropic diagonal; and three random

Table 3: Ten gauges on one sequence, one draw (the orbit is exact per draw; these absolute values are not comparable with the 400-draw medians of Table 1). Along a gauge orbit that moves the logits by 2.31×10^{-14} the customary orthogonality deviation moves by $255\times$ and the contrasts do not move at all; transformations *outside* the group move C_{op}, and nothing moves C_{route}.

	$\text{Orth}(A)$	C_{op}	C_{route}	$C_{\text{op}} \cdot C_{\text{route}}$	$\lvert\Delta\text{logit}\rvert_{\max}$
min over ten gauges	0.748	1.6241221	1.9874976	3.2279388	—
max over ten gauges	190.635	1.6241221	1.9874976	3.2279388	—
max/min	**254.89**	$1 + 3{\times}10^{-15}$	1.0	$1 + 3{\times}10^{-15}$	—
Control (one transformation inside the group, two outside)					
$A \mapsto AM$, $B \mapsto M^{-1}B$	—	1.624	1.987	3.228	2.31×10^{-14}
$A \mapsto AM$, B unchanged	—	1.803	1.987	3.584	21.9
A unchanged, $B \mapsto M^{-1}B$	—	1.372	1.987	2.726	442.0

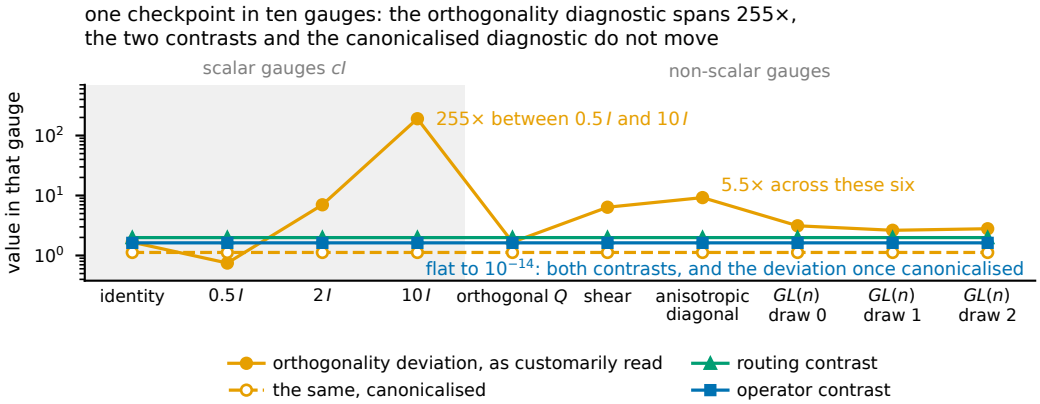

Figure 4: One trained checkpoint read in ten gauges $M \in GL(n)$: the orthogonality deviation spans $254.9\times$ across them ($5.5\times$ across the six non-scalar gauges alone), while the two contrast statistics, and the deviation once canonicalised, are constant to machine precision on the same axis. An in-group transformation leaves the logits unchanged to 2.31×10^{-14}.

$GL(n)$ draws. Measured spreads across the orbit: Orth ratio 254.892; C_{op} spread 5.11×10^{-15}; C_{route} spread exactly 0; product spread 9.77×10^{-15}.

The negative control applies three transformations that are *not* symmetries — $A \mapsto AM$ with B unchanged, $B \mapsto M^{-1}B$ with A unchanged, and a gauge followed by a scaling of B. The first two move C_{op} to 1.803 and 1.372 from 1.624, and move the logits by 21.9 and 442.0 in absolute terms. The instrument is therefore not blind; it is invariant on the group and responsive off it. The third is instructive: a gauge followed by scaling B by 1.3 moves the logits by 19.3 while leaving C_{op} at 1.624, because a global rescaling of B is absorbed identically by target and distractors and so cancels in a contrast. That is correct behaviour for a contrast statistic, and it is worth stating so the invariance is not over-read.

C THE MATCHED GATE FAMILIES

What is and is not new in Section 5. This is not load balancing (Shazeer et al., 2017; Fedus et al., 2022; Lample et al., 2019): load is first-order and can be balanced while the collision structure the ladder moves stays free. The quantity is mean square coherence (Barg et al., 2013), the interference variance of Massey & Mittelholzer (1993); average against worst-case coherence is Mixon et al.

(2011). We add a matched intervention in a trained model with a paired structural null, and no new statistic.

Table 4: The dose–response ladder of Section 5: accuracy on the shipped checkpoints with the trained gates replaced by each rung's constructed table. The last rung is the control the two-point version lacked.

$\mathbb{E}[b^2]$	shifts mixed	N=16	N=24	N=48	N=64
trained gates		0.9717	0.9378	0.8000	0.6909
0.007216	0	0.9255	0.8094	0.4954	0.3333
0.008867	4	0.8833	0.7559	0.4344	0.2968
0.010763	8	0.8286	0.6623	0.3600	0.2438
0.012353	12	0.7830	0.5975	0.2722	0.1818
0.010641	16	**0.8210**	**0.6644**	**0.3236**	**0.2114**

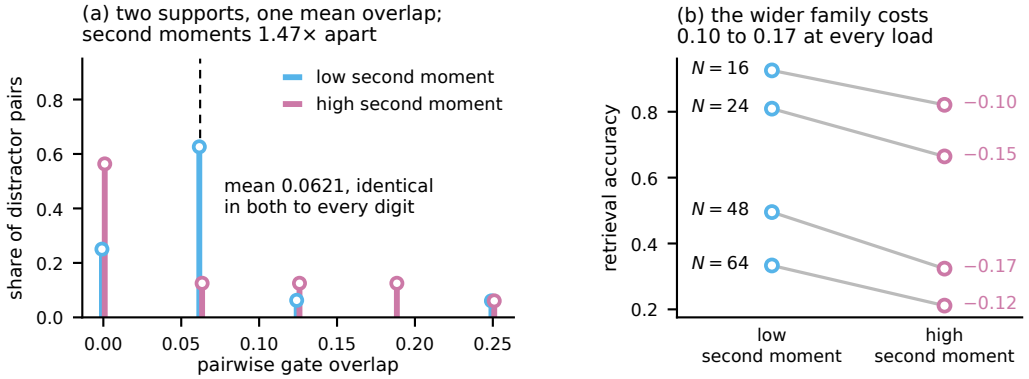

Figure 5: The matched-moment intervention: two support families with the same mean overlap (0.0621) and second moments 1.47× apart cost 0.10 to 0.17 accuracy at every load. (a) The overlap distributions of the two families over their true discrete support, the means coinciding exactly. (b) Paired accuracy under each family, one line per checkpoint.

D THE FRAME-POTENTIAL BOUND

Section 4 reports how much operator contrast is available above the random floor once the separation between values sharing a key is held fixed. This is that bound, with its assumptions.

Write each binding as $A(k)B(v)$ and normalise it to unit Frobenius norm, which is the cosine gauge of Section 2; the normalised writes are unit vectors in $\mathbb{R}^d$ with $d = n^2$. For K stored keys each carrying V values, the frame potential of the KV writes obeys $\left\| \sum_i x_i x_i^\top \right\|_F^2 = KV + \sum_{i \neq j} \cos^2_{ij} \geq (KV)^2/d$, the Welch bound in its second-moment form (Welch, 1974; Massey & Mittelholzer, 1993). Splitting the off-diagonal sum by whether the two writes share a key — $KV(V-1)$ same-key ordered pairs of mean c_s, $K(K-1)V^2$ different-key pairs of mean c_d^{kv} — and rearranging,

$$c_d^{\mathrm{kv}} \;\geq\; \frac{KV/d \;-\; 1 \;-\; (V-1)\, c_s}{(K-1)\, V}.$$

c_d^{kv} is the different-key mean over the stored-pair frame $\{A(k)B(v)\}$ alone. It is not the c_d of Eq. 3, which averages every prefix write — cross and query-side writes included, 60 to 63% of the population — against the full-vocabulary frame at the query key; that quantity is 0.0156 on every trained

checkpoint (`MARGIN_BOUND_20260917.json`), the spherical $1/d$, and the bound does not constrain it. The contrast of Section 4 is large when c_d^{kv} is small, so the bound is a ceiling on it. V here is the number of values a key can take, which is not the vocabulary: the official generator draws keys from the lower half of the vocabulary and values from the upper half, so at a 512-word vocabulary $V = 256$ and the key alphabet is 255. The two coherences are measured rather than assumed. On the random binding c_d^{kv} is 0.0160 to 0.0162, at the spherical value $1/d$, but c_s is 0.031 to 0.033, twice that: writes sharing a key share the factor $A(k)$ and are not independent. Training leaves c_d^{kv} where it was (0.0158 to 0.0162) and lowers c_s to 0.020–0.021, by 35% — the same-key separation of Eq. 5, which is the one angular quantity training moves. At $V = 256$, $d = 64$ and the random binding's own c_s, the bound leaves room of 5.4% at $K = 64$ rising to 11.2% at $K = 16$ below the measured c_d^{kv}; at the trained c_s, 4.1 to 6.6% (both emitted by `ceiling_fresh_kv.py --annotate`). Measured at the resampling optimum itself (`CEILING_FRESH_20260919.json`, a later pass), the same-key mean square cosine is 0.036 to 0.040, against 0.020 to 0.021 for the trained binding and 0.031 to 0.033 for the random one: 1.78 to 1.85 times the trained level and above the random level on all four checkpoints. The different-key mean rises with it, 0.018 to 0.019 against 0.0158 to 0.0162 trained, so the contrast is not bought by lowering c_d^{kv}. The 20 points of the accuracy-constrained front agree, at 1.5 to 2.2 times the trained level. Read on stored pairs alone, the optimum's cosine contrast is 0.81 to 0.90 times the random floor (the trained binding's, 0.99 to 1.03), while over all distractors it is 1.38 to 2.03 (`CEILING_FRESH_KV_20260920.json`, the same ascent rerun with its optimum saved; every overlap and ratio of the earlier file replays). The gain the search buys therefore sits in the cross and query-side writes, the 60 to 63% of the distractor population that the shift convention manufactures, not among stored pairs. The bound says it could not have sat there: at the optimum's own c_s it allows c_d^{kv} to fall 6 to 13 per cent below the random binding's, and c_d^{kv} rose instead. So the arrangement among stored pairs has the 5 to 11% of room stated above, the search does not take it, and what it raises is the angle of the artefact writes — the writes training separates by norm (Section 4) rather than by angle. The same-key overlap rises alongside and the accuracy falls; the inequality relates neither to the gain.

Four assumptions are load-bearing and none is hidden. Writes are taken at unit Frobenius norm. The contrast is read on means over the uniform set of all KV pairs, where the body reports medians of per-sequence ratios, and we do not supply the transfer between them. Pairs are assigned uniformly, where the generator samples without replacement within a sequence. And the bound is vacuous when $K \leq d$ and c_s is allowed near 1.

What the inequality does not give is the step the body most wants: that retrieval requires c_s near the random level. Squared inner products count anti-alignment as strongly as alignment, and vectors with a common component can still be separated by an argmax. The accuracy collapse reported in Section 4 is therefore an observation about the solution the search reached, not a theorem.

E PREREGISTRATION

The following were fixed in writing before the measurements of Sections 4–9 were run, and are reproduced in the supplementary artifact. They govern the contrast protocol — readout direction, distractor population, floor construction, gauge, stopping rule — and nothing else. The held-out and fresh-draw ceilings, the split of writes by type, the write-gate ablation, the dose–response ladder, the worst-overlap control, the exact pairwise computation and the frequency-weighted arrangement population are *post hoc*: each was designed after seeing the result it responds to, and each is reported with the protocol it actually used rather than under this registration.

Readout direction. $u = B(w^*)$, the answer token's row of the value codebook, alone. Not the argmax logit direction, not a marginalised direction, and on DeltaNet not the written residual $\tilde{v}_{s^*}$.

Distractor population. The sequence's stored keys, grouped by distinct key, so that repeated occurrences of one key form one distractor rather than several.

Floor. The same architecture at the same configuration and the same seed, taken at step 0, evaluated on identical draws so that the comparison is within-draw. Every learned tensor is at its initialisation: the two codebooks *and* the gate. This matters for reading C_{route}, which does not depend on A or B at all and could not move against a floor that randomised only the binding.

Gauge. The cosine normalisation of Section 2, reported alongside the raw value.

Stopping rule for the positive control (added after the first attribution). The committed registration fixes the certification gate (≥ 0.99) and the draw count; the rule that at most two further attributions may be tried, each justified before it is run, was adopted on 2026-09-15 after the pre-registered attribution had failed and before the second and third were run. It is why Section 9 reports a failure rather than a fourth definition. The registration's three predictions were: P1, the routed member's full contrast trained/floor near 1 — falsified (13 to 20×); P2 and P3 on the attribution rules, reported in Appendix G.

Estimator discipline. Contrast statistics are ratios with heavy tails. All reported values are medians over the stated number of draws, with the draw count stated. This rule was added after a single draw of a contrast ratio returned 1.63 on a checkpoint where 512 draws of the same quantity gave a median of 23.4 and a standard deviation of 256.4.

F FULL NUMBERS

Absolute contrast values (official generator, 400 draws). Trained, then floor: C_{full} 4708.6/239.2, 3328.5/233.1, 3210.9/235.5, 2951.6/224.4 for $N = 16, 24, 48, 64$; C_{op} 160.9/62.4, 154.1/61.7, 142.5/62.5, 127.9/62.0; C_{route} 22.63/3.74, 23.65/3.54, 26.96/3.54, 27.79/3.50.

Factorisation assumption (Section 2.4). $C_{\text{op}} \cdot C_{\text{route}}$ over C_{full}: 0.773, 1.095, 1.197, 1.204. Correlation between g_s^2 and m_s^2 across the distractor population: -0.038, 0.014, 0.036, 0.025.

Write norms by type (Section 4). Trained over floor, medians of $\|A(x_{s-1})B(x_s)\|$:

write	N=16	N=24	N=48	N=64
the queried pair	2.78	2.73	2.39	1.91
other true pairs	2.78	2.75	2.39	1.90
cross writes	0.23	0.24	0.45	0.47
query-side	0.64	0.64	0.57	0.48

Shared vocabulary (Section 10). The official generator draws keys from the lower half of the vocabulary and values from the upper half in its body (`key_vocab_size = vocab_size // 2`); no parameter reaches it, so the variant is one hunk applied to the official file's text at run time, drawing the $2N$ storage tokens together without replacement from the whole vocabulary — layout, power law, labels and every assertion unchanged, the diff recorded in `SHARED_VOCAB_20260919.json`. The four cells were retrained at the published settings (seed 17) and read with the packaged estimators executed verbatim; one draw of $2N$ replaces two of N, so the sequences are not token-for-token those of the disjoint cells. Disjoint / shared:

	N=16	N=24	N=48	N=64
true-pair over cross-write norm	12.16 / 0.99	11.37 / 1.01	5.28 / 0.98	4.07 / 1.01
operator contrast over floor	1.020 / 0.996	1.014 / 1.002	1.013 / 0.992	1.010 / 0.997
routing contrast over floor, true pairs	3.96 / 4.63	4.32 / 4.77	4.43 / 4.70	4.68 / 4.67
accuracy	0.970 / 0.784	0.936 / 0.640	0.800 / 0.338	0.693 / 0.237
final training loss, shared	0.0003	0.0019	1.13	2.35

Training still raises every write's norm over the floor (1.1 to 2.5×), by the same factor for true pairs, cross writes and query-side writes, so the gate of Section 4 is a reading of identity and nothing else; the operator's cross-key arrangement stays at its floor without it, and the routing contrast on true pairs does not move. Accuracy is not comparable across the two conventions: $N = 48$ and 64 did not converge in the budget, and nothing was retuned. The two codebook arms of Section 3 were then retrained under this convention at the two loads that converged (`SHARED_VOCAB_ARMS_20260920.json`, seed 17, same recipe; the routing arm was not). Disjoint / shared, accuracies and costs in points:

	full	scale only	codebooks frozen	magnitude cost	direction cost
N=16	0.972 / 0.784	0.952 / 0.742	0.823 / 0.687	12.9 / 5.5	2.0 / 4.2
N=24	0.938 / 0.640	0.913 / 0.590	0.710 / 0.541	20.4 / 4.9	2.5 / 5.0

Without the identity mark the magnitude cost falls from 13 and 20 points to about 5 and the direction cost rises from 2.0 and 2.5 to 4 to 5: most of what the write magnitudes were worth was the gate, and what remains is a scale training still needs; the directions are worth more once the gate cannot carry the type.

Paired resolution of the operator null (Section 4). Trained against floor on identical draws. Mean log ratio of the cosine-gauge contrast with its standard error, at 4096 paired draws (`OPERATOR_RESOLUTION_20260921.json`, declared before it ran; the same script at 512 reproduces the earlier pass to 10^{-9}): -0.0002 ± 0.0071, $+0.0001 \pm 0.0074$, $+0.0009 \pm 0.0079$, -0.0221 ± 0.0104, giving $|t|$ of 0.02, 0.01, 0.11, 2.12. Bootstrap 95% intervals on the median ratio: $[0.988, 1.010]$, $[0.990, 1.008]$, $[0.993, 1.009]$, $[0.994, 1.006]$. The smallest ratio separable from unity at two standard errors is 1.014, 1.015, 1.016, 1.021, against 1.037, 1.046, 1.078 and 1.054 at 512 draws and 4.1 to 6.6% of room in the frame bound. No cell separates; the one mean that moves at $|t| > 2$ ($N = 64$) moves *below* the floor, away from the room.

The evaluation-draw spread (Sections 5 and 10). The same untouched seed-17 checkpoint is read on four draws in this work: 32 sequences for the write-gate arms, 64 for the resampling ceiling, 256 for the permutation and arrangement tests and 500 held out. Across all four its accuracy spans N=16 0.0137; N=24 0.0083; N=48 0.0166; N=64 0.0107, so the widest is 0.017, at $N = 48$: 0.7998 on the 500-sequence draw against 0.8164 on the 32-sequence one (`PERMUTATION_ACCURACY_20260918.json`, `WRITE_GATE_ABLATION_20260916.json`). Over the three larger draws alone it is 0.014. The three-seed band of the trained model, all three read on the 500-sequence draw, is 0.0049: the draw is the larger nuisance, and both are far below every contrast the paper reports.

Signed margin (Section 4, 512 draws). Trained over floor: 28.3, 33.3, 36.6, 26.7. The margin is positive on every trained cell and near zero on every floor cell, which is the sanity condition that the DeltaNet control of Section 9 fails.

The matched families (Section 5). Both families place uniform mass on 4 of 16 slots and use 16 distinct supports. Low-var uses grid lines, high-var uses cyclic shifts. Mean pairwise overlap 0.0621331 in both, to all printed digits. Second moment 0.0072162 against 0.0106409. The largest overlap between distinct supports is $2/16$ on the grid family and $3/16$ on the cyclic one; along the ladder it is $2/16$ at rung 0 and $3/16$ at rungs 4, 8, 12 and 16, so the mean-matched pair 4 against 12 is matched on it and the pair 0 against 16 is not. The number of distinct supports runs 16, 15, 14, 13, 16; the pair 4 against 12 differs in full-collision symbol pairs by 23%. One pair is not ordered by the second moment: at $N = 16$, 48 and 64 the eight-shift rung is worse than the sixteen-shift rung on the mean overlap, the second moment and the number of distinct supports, and is nonetheless 0.8, 3.6 and 3.2 points more accurate ($N = 24$ is perfectly ordered). The rank correlation of -0.90 rather than -1 is that pair, and we do not know what orders it. A separate arm caps the maximum directly — grid lines against every other cyclic shift, both topping out at $2/16$ — and moves accuracy by 0.1255 to 0.1943 as the second moment rises from 0.007216 to 0.011619; it reaches that second moment with 8 distinct supports instead of 16, so it is supporting evidence rather than a clean control. The maximum over all symbol pairs is $4/16$ on every rung, because 512 symbols reuse 16 support rows; that statistic cannot separate any two rungs and is not the one quoted here. A global relabelling of the slots, which changes no overlap, moves accuracy by exactly 0 on all four checkpoints — a check on the harness rather than a null, since slot permutation is an exact symmetry of the model. The 200 per-symbol slot permutations of Table 6 (the concentration-matched null of Section 6) move accuracy by -0.005, -0.009, -0.023, -0.031 on average — three to twenty times smaller than the matched contrast, in the same direction on every draw at $N \geq 48$ — and their load is free to rise. The 50 load-matched nulls of Section 6 move it by -0.001, -0.001, -0.006, -0.011, with 15, 16, 2 and 0 of the 50 scoring at or above the trained router (Table 5); most of the permutation cost is load. The accuracy column of Table 6 is the permutation mean.

Table 5: Accuracy under the two arrangement nulls, all draws, a later pass: mean cost and [min, max] over the 200 per-symbol permutations and over the 50 load-matched nulls, on the draw of Table 6 and on the training script's held-out draw that the relation uses (file `PERMUTATION_ACCURACY_20260918.json`). An earlier text quoted one permutation (seed 18): 0.0002, 0.0236, 0.0324, 0.0409; at $N = 24$ that draw lies outside the range of the 200.

draw	N	trained	200 permutations, mean [min, max]	50 load-matched, mean [min, max]
Table 6	16	0.9717	0.0051 [−0.0032, 0.0144]	0.0014 [−0.0029, 0.0115]
	24	0.9378	0.0086 [−0.0016, 0.0207]	0.0011 [−0.0070, 0.0114]
	48	0.8000	0.0230 [0.0085, 0.0487]	0.0060 [−0.0009, 0.0125]
	64	0.6909	0.0306 [0.0143, 0.0526]	0.0109 [0.0024, 0.0197]
held-out	16	0.9699	0.0049 [−0.0017, 0.0119]	0.0021 [−0.0034, 0.0058]
	24	0.9363	0.0101 [0.0012, 0.0231]	0.0033 [−0.0029, 0.0082]
	48	0.7998	0.0249 [0.0125, 0.0527]	0.0076 [0.0017, 0.0125]
	64	0.6931	0.0319 [0.0162, 0.0528]	0.0130 [0.0062, 0.0190]

The reference has a closed form. For two symbols whose slot masses are permuted independently and uniformly, $\mathbb{E}[b^2] = q_a q_b/M + (1 - q_a)(1 - q_b)/(M(M - 1))$ with $q = \sum_i g_i^2$; against the 200-draw empirical mean it agrees to better than 0.5% at every load, weighted or not, so no sampling is required at any N.

What the permutation does not preserve is the aggregate. Since $\sum_{ij} p_i p_j \langle g_i, g_j \rangle = \| \sum_i p_i g_i \|^2$ for any weights p, the frequency-weighted slot load is a function of between-symbol structure, and a per-symbol permutation is free to change it. Measured against the same 200 permutations, the trained router's $\| \sum_i p_i g_i \|^2$ is 0.0625 to 0.0628 against a permutation mean of 0.0640 to 0.0654: lower by 1.9% at $N = 16$ rising to 4.3% at $N = 64$. The arrangement gap of Section 6 is therefore measured against gates matched per symbol and not on load, and two controls bracket the load's share of it. Centring each gate row on the frequency-weighted mean before the coherence is taken removes the load from the statistic; against the same 200 permutations the trained router then sits 4.0 to 4.9% below (ratios 0.960, 0.954, 0.951, 0.952 at $N = 16/24/48/64$; rank 0 of 200; `SPECTRAL_DESCRIPTORS_20260917.json`). And a null annealed from the per-symbol permutation until the weighted load matches the trained router's to a relative 10^{-4}, 50 draws per checkpoint, holds concentration and load together: the trained router sits 0.2 to 0.6% below its mean (ratios 0.994, 0.994, 0.994, 0.998), rank 0 of 50, $z = -2.3$ to -4.0 (`LOAD_MATCHED_NULL_20260917.json`); the annealing destroys some arrangement while it matches, so that reading is conservative.

Table 6: The trained router against 200 concentration-matched permutations of itself, each row read against its own permutation: rank 0 of 200 on all four (one-sided $p = 0.005$, the smallest this design can report), z in units of the permutation distribution's own spread. At $N = 16$ the permutations land on the analytic reference for independent uniform supports, $\mathbb{E}[b^2] = 0.006250$. The last two columns weight each symbol pair by how often the generator presents the symbols as keys, from a separate draw of 200 permutations; that is the reading quoted in Section 6. The accuracy cost is the mean over the 200 permutations (ranges and the load-matched cost: Table 5).

pairs N	trained $\mathbb{E}[b^2]$	permuted $\mathbb{E}[b^2]$	ratio	z	accuracy cost	weighted ratio	weighted z
16	0.006133	0.006277	**0.977**	−2.55	−0.0051	**0.956**	−2.29
24	0.007159	0.007374	**0.971**	−2.93	−0.0086	**0.947**	−2.96
48	0.010304	0.010722	**0.961**	−3.05	−0.0230	**0.942**	−3.03
64	0.013518	0.014094	**0.959**	−2.77	−0.0306	**0.941**	−2.74

The concentration-matched null (Section 6). Table 6 reports 200 independent per-symbol permutations per checkpoint, with bootstrap intervals from 2000 resamples. The superseded sampled estimator instead used 2000 paired draws and recorded the ratio, the t statistic and the fraction of draws won; it gives a mean log ratio of -0.0184 with standard error 0.0191 on the $N = 24$

checkpoint. The exact computation over all $V(V-1)$ ordered pairs, against 200 permutations, gives a ratio of 0.971 at rank 0 of 200, $z = -2.9$ ($p = 0.005$), on the same checkpoint; the exact estimator's standard error in log ratio is 0.0008 against the sampled estimator's 0.0191. The gap between those two lines is the finding of Section 6. Constructed ensembles at a fixed effective slot count $\exp H = 3.596$ and a fixed participation ratio 3.333 span C_{route} from 1.000 (aligned) through 13.54 (spread) to 19.27 (collision-avoiding).

Seed replication (Reproducibility Statement). Answer accuracy per cell, retrained from scratch at each seed on the official generator, 10,000 steps.

pairs N	seed 17	seed 29	seed 41	spread
16	0.9717	0.9728	0.9735	0.0018
24	0.9378	0.9356	0.9401	0.0045
48	0.8000	0.7992	0.8040	0.0049
64	0.6909	0.6948	0.6946	0.0039

Seed 17 is read on the 256-sequence draw and seeds 29 and 41 on the 500-sequence one; all three on the latter give the same maximum, 0.0049.

Per-cell contrasts on the two further seeds, which the Reproducibility Statement quotes as ranges. C_{route} and the cosine-gauge C_{op} are ratios to the paired floor; the arrangement column is the trained router over its own permutation.

	C_{route}		C_{op} (cosine)		arrangement	
pairs N	seed 29	seed 41	seed 29	seed 41	seed 29	seed 41
16	6.99	6.87	0.943	0.989	0.975	0.975
24	7.75	7.64	1.000	1.015	0.971	0.970
48	8.44	8.81	1.002	1.001	0.959	0.960
64	8.56	8.91	0.999	1.018	0.958	0.960

The superseded power curve (Section 6). Blending the trained gates toward the structured family with weight α and rerunning the sampled estimator at 2000 paired draws gave $t = -0.96, -0.96, -0.96, -0.97, -1.06, -2.12, -8.47$ for $\alpha \in \{0, 0.02, 0.05, 0.1, 0.2, 0.4, 1.0\}$. We recorded that curve as the power of the null, and it is not: it is the power of an estimator we have since replaced. The exact computation resolves at rank 0 of 200 ($z = -2.3$ to -3.0) an effect this curve placed below its own detection threshold, so the curve bounds the sampling, not the phenomenon. It is kept here because the comparison between the two lines is the point of Section 6.

Operator ceiling (Section 4). Adam on the codebooks against the cosine-gauge contrast, 400 steps at $\text{lr} = 0.05$, run twice: from the random binding and from the trained one. Optimised on one draw set and read on a disjoint one, with the floor and the trained reading taken on the reading set. Optimising the mean and reporting the median, so the reported value is not the quantity being maximised. The result is a lower bound on what is achievable, which is the direction that matters here.

pairs N	floor	trained	same-draw	held out	accuracy at the optimum
16	65.53	65.60	1.08×	0.99×	0.958 → 0.680
24	63.92	64.44	1.17×	1.00×	0.945 → 0.646
48	63.90	63.98	1.53×	1.01×	0.803 → 0.503
64	64.20	64.33	1.76×	1.00×	0.702 → 0.405

Starting the same fixed-set search from the random binding costs 43.2 to 49.9 accuracy points instead of 27.8 to 30.0. The resampling search (a fresh draw at every step, read on a disjoint set) reaches 1.38, 2.03, 2.02 and 1.96× the floor at a cost of 0.217, 0.266, 0.263 and 0.256.

The accuracy-constrained search (Section 10). The first version of this search did not constrain anything: its penalty pushed the optimised codebooks into the frozen model with a non-differentiable

copy, so the cross-entropy it added was a constant and every λ arm descended pure contrast (Appendix H). Re-run with the penalty differentiable (`CEILING_PARETO_20260919.json`, the assertion that its gradient is non-zero checked at every arm), the largest contrast held within two accuracy points is $1.03\times$ the floor, and at the two rates where the unconstrained search reaches 1.2 to $2.0\times$ no λ satisfies the constraint at any checkpoint at 0.05, and at 0.02 only at $N = 16$, where the two satisfying arms lift nothing (1.004 at most): the front is a cliff rather than a curve. Against the $\lambda = 0$ arm of the same checkpoint, 33 of 36 penalised arms recover accuracy (by 0.001 to 0.122), but only 3 raise the contrast ratio, and those same 3 (all at $\mathrm{lr} = 0.005$) are the only arms that improve on both, by at most 0.007 accuracy and $0.02\times$ the floor: the penalty buys accuracy back by giving up the contrast the unconstrained search had found. The figures below are from the superseded run and are kept because the text above reports what the correction changed. Its protocol was the resampling objective plus λ times the task cross-entropy, 300 steps: at $\mathrm{lr} = 0.02$ and 0.005 from the trained codebooks with $\lambda \in \{0, 10^2, 10^3, 10^4\}$, at $\mathrm{lr} = 0.05$ from the random binding with $\lambda \in \{0, 10, 10^2, 10^3\}$. At $\mathrm{lr} = 0.005$ nothing moves: contrast 0.99–$1.09\times$ the floor, accuracy within 0.003 to 0.033 of the trained value. At $\mathrm{lr} = 0.05$ from the random binding the contrast reaches 1.08–$2.05\times$ at accuracies of 0.51–0.90. At $\mathrm{lr} = 0.02$, the rate that moves the contrast from the trained point, contrast ratio and accuracy are, for the four weights in order:

pairs N	trained	$\lambda = 0$	$\lambda = 10^2$	$\lambda = 10^3$	$\lambda = 10^4$
16	0.958	$1.07\times$ /0.944	$1.03\times$ /0.948	$1.06\times$ /0.939	$1.08\times$ /0.931
24	0.945	$1.24\times$ /0.889	$1.12\times$ /0.885	$1.22\times$ /0.879	$1.14\times$ /0.887
48	0.803	$1.68\times$ /0.677	$1.83\times$ /0.677	$2.02\times$ /0.705	$1.79\times$ /0.667
64	0.702	$1.96\times$ /0.572	$1.97\times$ /0.576	$1.98\times$ /0.574	$1.73\times$ /0.499

The weight on the task loss has no ordered effect at any rate; what the contrast gains, the task pays, at every rate that moves it.

The task-loss-only control (Section 10). At each rate of the constrained search, the codebooks were also descended on the task loss alone from the trained checkpoint, same steps, no contrast term (`TASK_ONLY_CONTROL_20260919.json`). Accuracy cost in the task-only descent / in the constrained search's unpenalised arm at the same rate:

	lr 0.05	lr 0.02	lr 0.005
N=16	0.078 / 0.064	0.021 / 0.018	−0.005 / 0.001
N=24	0.092 / 0.134	0.039 / 0.062	0.011 / 0.005
N=48	0.079 / 0.191	0.026 / 0.109	−0.007 / 0.017
N=64	0.077 / 0.173	0.015 / 0.131	0.000 / 0.026

At $N = 16$ the descent alone costs more than the search at both rates that cost anything, so none of the search's cost there is attributable to its objective; at the other loads the descent accounts for part of it; at the smallest rate for nothing at $N = 48$ and $N = 64$, while at $N = 24$ it again costs more than the search (0.011 against 0.005). The resampling search of Section 4 (21.7 to 26.6 points) is a different protocol, 400 steps at rate 0.05 with a fresh draw per step, and has no such control.

Withheld learning (Section 3). The three trained arms retrained at seeds 29 and 41, same cells, same budget; accuracy at seeds $17/29/41$. The nothing-trained arm is evaluated at each seed's initialisation. The frozen router is the initialiser's own draw, unit-variance Gaussian logits over $M = 16$ slots at temperature 1, redrawn per cell: across the twelve router-frozen checkpoints the per-symbol effective slot count $\exp H$ averages 10.4 to 10.7 and the mean top probability 0.24 to 0.25, against 3.7 and 0.57 once the router trains on the codebook-frozen arm at $N = 16$. The routing cost is the distance from a diffuse random routing at that logit scale; a frozen router drawn sharper or flatter would cost a different amount, and the scale was not varied.

trained	N=16	N=24	N=48	N=64
codebook scale only	.952/.955/.957	.913/.907/.910	.743/.745/.745	.646/.644/.637
codebooks frozen	.823/.812/.813	.710/.701/.697	.471/.490/.471	.362/.375/.361
router frozen	.594/.600/.583	.425/.459/.423	.245/.265/.267	.217/.199/.188
nothing trained	.223/.184/.180	.140/.119/.114	.065/.049/.053	.047/.036/.036

Across the three seeds the cost of a frozen router is 37 to 55 points, of frozen write magnitudes 13 to 28, of frozen directions 2 to 6; the largest seed spread on any trained arm and load is 0.036 (the nothing-trained arm spreads by 0.043 at N=16).

A frozen random hard hash (Section 3). The frozen router above is the initialiser's draw, diffuse over about 10.5 effective slots. Run after a later internal pass and declared before it ran (`FROZEN_HASH_ROUTER_20260920.json`): the gate table set to a random balanced assignment of the 512 symbols to the 16 slots, 32 per slot, one-hot at logit 30 and frozen; the codebooks trained under the recipe of Section 3 at seeds 17, 29 and 41, each seed drawing its own assignment. Accuracy and cost in points beside the diffuse frozen router, seed order 17/29/41:

	full	hard hash	cost	diffuse	cost
N=16	0.972/0.973/0.974	0.969/0.972/0.972	0.2/0.1/0.1	0.594/0.600/0.583	38/37/39
N=24	0.938/0.936/0.940	0.935/0.934/0.935	0.2/0.1/0.5	0.425/0.459/0.423	51/48/52
N=48	0.800/0.799/0.804	0.786/0.788/0.790	1.4/1.1/1.4	0.245/0.265/0.267	55/53/54
N=64	0.691/0.695/0.695	0.688/0.687/0.690	0.3/0.8/0.4	0.217/0.199/0.188	47/50/51

The hash costs 0.1 to 1.4 points where the diffuse router costs 37 to 55: the price of the diffuse router is its diffuseness, and a concentrated, balanced assignment recovers all but a cost of the order of the arrangement channel of Section 6 without learning. Three seeds, each with its own random assignment; the trained model's seed spread is 0.005. Per load the hash costs 0.08 to 0.25 at $N = 16$, 0.14 to 0.54 at 24, 1.11 to 1.43 at 48 and 0.28 to 0.80 at 64, against a full-model seed spread of 0.18 to 0.48 points: only $N = 48$ is resolved above that spread; the other three loads bound the residual rather than measure it. Read with the estimator of Section 2 on the same draws (`HASH_ROUTER_CONTRAST_20260921.json`), the hash's routing contrast is 15.9 to 17.0 against a floor of 3.1 to 3.8, where the trained routers read 22.1 to 28.9; its mean square coherence is 0.0607 on every cell, the same-slot probability of a balanced one-hot table, against 0.0059 to 0.0135 for the trained routers.

The learning-rate sweep (Section 3). Every withheld arm had been trained at the rate the full model was tuned at, so a cost could include a mismatched step size. The sweep was declared before it ran (`FROZEN_GEOMETRY_LR_SWEEP_20260919.json`, seed 17): rates $\{3.16 \times 10^{-3}, 10^{-2}, 3.16 \times 10^{-2}, 10^{-1}\}$ for the four trained arms and the full model alike, each cell choosing its rate on a held-out draw disjoint from the training and test draws and reporting on the test draw. The pre-declared contingency dropped the last block, 3.16×10^{-3}, for every arm and load when the projected finish passed eight hours, so the surviving grid is $\{10^{-2}, 3.16 \times 10^{-2}, 10^{-1}\}$. Nothing diverged; the selection draw sits within 0.011 of the test draw on all 48 trained cells; at the incumbent rate the twelve withheld cells and the nothing-trained control replay their published accuracies exactly and the full arm within 0.0022 of the accuracies quoted for the trained model. 12 of 16 cells moved off the incumbent rate: the frozen-router arm to 10^{-1} at every load, the full model at $N = 24$ and 64. Costs in points at the incumbent rate and at the best of the grid on both sides:

cost	N=16	N=24	N=48	N=64
routing	37.6 → 37.0	51.1 → 49.5	55.7 → 55.5	47.6 → 48.7
write magnitudes	12.9 → 13.3	20.4 → 20.6	27.2 → 27.3	28.4 → 27.4
codebook directions	1.8 → 1.5	2.3 → 2.7	5.9 → 5.8	4.8 → 6.4

No cost moves by more than 1.6 points at any load, the shifts run both ways across loads, and the order routing > magnitudes > directions, costs against initialisation, holds at every load at both rates (against the hash the learned share of routing sits below the direction cost). Two residues: at $N = 16$, 24 and 64 the incumbent 10^{-2} is bracketed only from above, and one cell (codebooks frozen, $N = 48$) chose the lowest surviving rate. The main text keeps its three-seed ranges; the sweep is one seed.

The relation (Sections 2.4 and 8). On the twelve hash-router cells (`MARGIN_BOUND_20260921_hash.json`, run after the arm), Eq. 4 lands 0.011 to 0.059 below the measurement, the union bound below it on every cell; g_s is 0 or 1 there, the case the Gaussian step is least built for. Table 7 gives the ranges by group. The three scalars of Eq. 5 are

ratios of means, while the prediction needs the mean over queries of the per-query ratio; the two part by Jensen's inequality wherever routing weights and write norms vary across queries, which is why their product misses by up to 0.34 where the descriptors do not.

Table 7: The relation against measurement, by group. Accuracy is the model's argmax on the 500-sequence test draw; the prediction uses only the inputs of Eqs. 3–5; "var" is the measured over the predicted crosstalk variance (assumption ii, median over queries); "bound" is the union lower bound of Eq. 4. Ranges over the four loads $N = 16/24/48/64$.

cells		accuracy	predicted	bound	var
trained, $(M, n) = (16, 8)$, 3 seeds	12	0.69–0.97	0.64–0.96	0.19–0.89	0.95–0.97
codebook scale only, 3 seeds	12	0.64–0.96	0.58–0.93	0.18–0.84	0.94–0.97
codebooks frozen, 3 seeds	12	0.36–0.82	0.32–0.78	0.05–0.60	0.96–0.98
router frozen, 3 seeds	12	0.19–0.60	0.18–0.55	0.00–0.09	0.94–0.99
nothing trained, 5 coordinates	28	0.01–0.73	0.01–0.72	0.00–0.35	0.87–1.00
$M \in \{8, 32\}$, $n \in \{4, 16\}$	16	0.17–1.00	0.14–1.00	0.01–1.00	0.94–1.00
ladder, five rungs	20	0.19–0.92	0.18–0.89	0.00–0.69	0.97–1.00

Per cell at the paper's coordinate, seed 17: measured accuracy / prediction / union bound, with the crosstalk variance ratio in brackets. The remaining 84 cells, with the cross-term factor, the effective interferer count $P_N^2 / \sum_s (g_s^2 \|W_s\|^2)^2$ (medians 3 to 6 on the trained model, 18 to 50 with the router frozen over the three seeds, up to 149 at initialisation) and the candidate-set check, are in the shipped `MARGIN_BOUND_20260917.json`.

	N=16	N=24	N=48	N=64
everything	.970/.955/.87 [.97]	.936/.916/.76 [.97]	.800/.753/.36 [.97]	.693/.639/.19 [.96]
codebook scale	.952/.920/.82 [.96]	.913/.873/.71 [.97]	.743/.672/.32 [.95]	.646/.576/.18 [.95]
codebooks frozen	.823/.773/.58 [.98]	.710/.663/.41 [.97]	.471/.425/.13 [.97]	.362/.321/.05 [.96]
router frozen	.594/.543/.08 [.96]	.425/.399/.03 [.98]	.245/.229/.01 [.99]	.217/.200/.01 [.98]
nothing	.223/.200/.03 [.98]	.140/.133/.01 [1.00]	.065/.063/.00 [.98]	.047/.047/.00 [.96]

Three other predictors are reported for calibration and none is used: the Gaussian model with each query's *measured* crosstalk moments misses by up to 0.21 at the highest loads (the tail is heavier than Gaussian where a few interferers dominate); one signal-to-noise ratio per cell, from the geometric mean over queries, misses by up to 0.55 (averaging descriptors instead of predictions); and a per-competitor coherence in place of the single c_d changes the prediction by less than 0.01.

The relation with the covariance the model implies. Assumption (i) takes e_w independent of the true-logit term e_{w^*}. The isotropic model behind Eq. 3 does not: it gives $\mathrm{Cov}(e_{w^*}, e_w) = P_N c_d \langle W^*, W(k,w) \rangle = \rho_w \sigma_*^2$. Keeping that covariance and conditioning on e^* replaces, in Eq. 4, e^* by $(1 - \rho_w) e^*$ and σ_w by $\sqrt{\sigma_w^2 - \rho_w^2 \sigma_*^2}$, and the union-bound denominator by $\sqrt{P_N c_d \|W^* - W(k,w)\|^2}$. Re-evaluated on the same 112 cells (`MARGIN_BOUND_20260918_cov.json`, a later pass; no conditional variance needed its floor), the prediction moves by at most 0.043: it lands within 0.054 of the measurement (against 0.073 without the covariance), mean -0.018, above the measurement on 10 cells by at most 0.013 (against 11), with the union bound above it on 2 (against 2, by at most 0.006); the rank correlation with measurement is 0.999 (against 0.999) and the order of the five arms holds at all 12 seed-load groups, as before. The ordering statements of Section 8 are unchanged by the correction.

Table 8: Across slot count M and binding dimension n the cosine-gauge operator contrast stays at its floor (bold) while the routing advantage grows with M, and n alone moves accuracy. Seed 17; C_{op} over the paired floor; arrangement: unweighted mean square coherence over the router's own permutations, rank among 200 and z.

sweep	M	n	N=16	24	48	64	C_{route} (×)	C_{op}	arrangement	rank, z
M (n=8)	8	8	.888	.786	.501	.376	4.6–5.5	**0.99–1.01**	0.972–0.983	0–23, −1.1 to −1.7
	16	8	.972	.938	.800	.691	6.1–7.9	**1.01–1.02**	0.959–0.977	0, −2.6 to −3.1
	32	8	.993	.984	.955	.915	11–14	**0.98–1.01**	0.888–0.918	0, −3.1 to −5.2
n (M=16)	16	4	.652	.525	.257	.166	8.1–9.0	**0.99–1.03**	0.940–0.947	0, −2.1 to −2.4
	16	16	.997	.994	.981	.978	3.7–6.4	**0.99–1.00**	0.951–0.976	0–3, −1.7 to −3.1

The scaling of the required contrast (Section 6). At fixed absolute accuracy the required routing contrast scales as $(N-1)^{1.026}$ with $R^2 = 0.9905$ over the $4\times$ range of pair counts.

Condition numbers as a component fingerprint. A condition number is often read as a quality: a well-conditioned memory retrieves cleanly. Here it is read instead as a record of what each learned component does to the state a query decodes from. At a query t the state is $R_t = \sum_{s\le t} g_s W_s$; the Gram of its writes over the prefix, $\sum_s a_s^2 \hat{W}_s \hat{W}_s^\top$ with unit writes $\hat{W}_s$, is conditioned over its nonzero spectrum (a prefix of t writes spans at most $\min(t, d)$ directions) with the three components switched on independently: *geometry* is the write directions from the trained codebooks (without it, the directions of the same seed's initialisation, the paper's random-binding floor), the *gate* is the write norm $\|W_s\|$, *routing* is g_s; the norm is Frobenius because Eq. 1 reads writes through the Frobenius inner product. Medians over queries, $\log_{10}$:

	N	geometry	gate	routing	geom.+gate	geom.+routing	all	initialisation
trained	16	2.1	3.8	3.8	3.8	3.8	5.3	2.6
trained	24	2.3	3.9	4.2	3.9	4.2	5.6	2.7
trained	48	1.3	2.0	3.6	2.0	3.6	4.4	1.6
trained	64	1.0	1.6	3.6	1.6	3.6	4.3	1.3
codebooks frozen	16	2.1	2.2	9.8†	2.2	9.8†	9.9†	2.6
codebooks frozen	24	2.3	2.3	9.7†	2.3	9.7†	9.7†	2.7
codebooks frozen	48	1.3	1.3	4.6	1.3	4.6	4.6	1.6
codebooks frozen	64	1.0	1.1	3.5	1.1	3.5	3.5	1.3
codebook scale only	16	2.1	5.3	5.0	5.3	5.0	7.7	2.6
codebook scale only	24	2.3	5.4	5.5	5.4	5.5	8.0†	2.7
codebook scale only	48	1.3	3.9	4.7	3.9	4.7	6.8	1.6
codebook scale only	64	1.0	2.2	3.9	2.2	3.9	5.9	1.3
router frozen	16	2.1	3.3	2.4	3.2	2.4	3.6	2.6
router frozen	24	2.2	3.3	2.5	3.3	2.5	3.6	2.7
router frozen	48	1.3	2.3	1.5	2.3	1.5	2.6	1.6
router frozen	64	1.0	1.9	1.2	1.9	1.2	2.2	1.3

The geometry column does not move under training, and the gate and routing columns read the same whether the directions are the trained ones or the initialisation's, to the first decimal at every load and arm: in conditioning the trained directions are indistinguishable from random ones. The write gate multiplies the condition number by 4 to 48 and routing by 43 to 340, and the model as it reads sits nearly three orders of magnitude above its initialisation — at the highest accuracy in the paper. With the codebooks frozen the router compensates by concentrating harder still: at $N = 16$ the stable rank of the read state falls from 15.3 under geometry alone to 2.0 of a rank-50 prefix, the state being close to rank-deficient in every direction but the target's slot. We give that one as a stable rank because its condition number is past the resolution of the statistic (below), and with the router frozen the gate alone accounts for the rise. So conditioning is a faithful fingerprint of which component acts, and it moves against retrieval, not with it. Two things follow. A diagnostic that reads the geometry alone cannot register what training changed in this model, because here the components it omits are the

ones that move (on DeltaNet the geometry column itself moves across arms, Appendix G). And no condition number carries a claim about accuracy in this paper: the relation of Section 2.4 is built from the descriptors instead.

What a condition number can resolve, and what we report beside it. The Gram is formed explicitly as $X^\top X$ and κ is its own condition number, $\lambda_{\max}/\lambda_{\min}$, the square of $\kappa(X)$. Its eigenvalues carry a rounding error of order $\epsilon\lambda_{\max}$, so the arithmetic resolves κ itself up to $1/\epsilon$; we draw the line at the stricter $\tau = \sqrt{\epsilon}$, that is κ below $1/\sqrt{\epsilon} \approx 10^{7.8}$, where $\kappa(X) = \sqrt{\kappa}$ would still be a measurement, because that is the line the re-run test below vindicates. We report, at every cell of both memories, the stable rank $\|X\|_F^2/\|X\|_2^2$ — bounded by the true rank, insensitive to the smallest singular values, and readable where κ is not — together with the fraction of the spectrum below $\tau = \sqrt{\epsilon}$ (`KAPPA_COMPONENTS_20260919.json` and its DeltaNet counterparts, run after the review). Seven of the 112 routed-memory cells sit past that line, all of them in the frozen-codebook arms at $N = 16$ and $N = 24$: the routing and all-three cells of the codebooks-frozen rows and the all-three cell of codebook-scale-only at $N = 24$, marked † in the table; no ratio or ordering in this paper rests on them. The tolerance is not a theoretical caution: re-running the identical arithmetic on the same machine with the same seeds, the 385 DeltaNet cells whose spectrum lies wholly above it came back bit-identical, while 9 of the 95 below it moved in the fourth significant digit, the worst where 48% of the spectrum is unresolved, never in the decimal printed. The field predicts which condition numbers are irreproducible, with no false positives. The ordering the body states survives the change of statistic and sharpens under it: the geometry column's stable rank is constant across the four arms to 0.1%, against 5.6% for its condition number, while the gate lowers it by 1.4 to 1.6× and the routing by 5.0 to 5.4× — a separation κ does not make at $N = 16$, where it ranks the two the other way round.

Arrangement (Section 6). 200 independent per-symbol permutations per checkpoint, $\mathbb{E}[b^2]$ computed exactly over all $V(V-1)$ ordered pairs of distinct symbols, bootstrap intervals from 2000 resamples. The permutations land at $\mathbb{E}[b^2] = 0.006277$ to 0.014094 against an analytic reference of 0.006250 for independent uniform supports at $N = 16$; that agreement is the check that the null is the null.

G THE DELTANET SCOPE TEST

We take the convention of the official reference, whose state is d_k by d_v and whose read is $S_t^\top q_t$; the written residual of Section 9 is $\tilde{v}_s = \beta_s(v_s - S_{s-1}^\top k_s)$ in it. The recurrence $S_t = P_t S_{t-1} + \beta_t k_t v_t^\top$ with $P_t = I - \beta_t k_t k_t^\top$ expands, from $S_0 = 0$, to $S_t^\top q_t = \sum_{s\le t} \beta_s v_s k_s^\top P_{s+1} \cdots P_t q_t$, the transitions needing no transpose because each is symmetric. A coefficient for each stored value therefore exists; it depends on every key after s, which is why a static per-position grouping is not it, and why the failure below is a statement about our instrument rather than about the concept.

The forgetting axis. The product $P_{s+1} \cdots P_t$ has a Lyapunov spectrum, estimated by QR accumulation over 64 generator sequences after a burn-in of 2,000 tokens, driven by each checkpoint's own trained key projection. Its top exponent is -0.0002 to -0.0007 per token on the four checkpoints (both heads), against -0.0035 to -0.0051 for isotropic keys; its spread $\lambda_1 - \lambda_d$ is 0.06 to 0.41. Three hypotheses the spectrum raised were each tested once. *Protected subspace*: the trained keys hit the slowest direction at 0.01 to 0.13 of the isotropic rate — confirmed. *Erasure channel*: the four fastest-forgetting directions carry 4.4 to 13.8 times their dimension share of key energy — confirmed. *Correlated writes* as the source of the stationary state's conditioning ($\kappa(S_\infty) = 325$ to 2,020 against 6 to 79 for independent writes): permuting the value stream changes κ by a factor of 0.29 to 3.0 and the key stream by 0.34 to 3.8, in both directions, and removing the common value component leaves it unchanged — not supported. The stationary state is in any case a regime the task never reaches: at $\lambda_1 = -0.0002$ to -0.0007 the initial state is forgotten over 1,500 to 6,300 tokens, and every sequence here is 512 tokens or fewer, so the state a query decodes from is transient, which is what the decoding analysis below reads. For memories with a diagonal, input-dependent transition the exponent per channel is exactly the mean log decay, $\lambda_c = \mathbb{E}_t[\log \alpha_{t,c}]$, read from the trained model on generator tokens: on the $N = 8$ cell (GLA, Mamba-2 and DeltaNet certified at ≥ 0.99; RetNet at 0.9855, below the gate, kept for its fixed decays), GLA (Yang et al., 2024a)

has $\lambda_{\min} = -1.24$ (a channel that forgets within a token) and λ_1 at zero, Mamba-2 (Dao & Gu, 2024) -0.055 to -0.0013, RetNet (Sun et al., 2023) -0.032 to -0.016 by construction, DeltaNet -0.41 to -0.0003, and this paper's memory 0 in every channel. GLA, Mamba-2 and RetNet were trained at $\mathrm{lr} = 3 \times 10^{-3}$ (DeltaNet 10^{-3}), state matched to DeltaNet. A sweep declared before it ran (`BASELINE_LR_SWEEP_20260920.json`: rates 10^{-3}, 3×10^{-3}, 10^{-2}, same cell, same script, seed 17, on a second GPU) reads:

	rate	accuracy	certified	λ_1	$\lambda_{\min}$
GLA	10^{-3}	0.9942	yes	0	-1.6
	3×10^{-3}	0.9961	yes	0	-1.3
	10^{-2}	0.9958	yes	0	-3.1
Mamba-2	10^{-3}	0.9916	yes	-0.0014	-0.016
	3×10^{-3}	0.9970	yes	-0.001	-0.07
	10^{-2}	0.9952	yes	-0.003	-0.0033
RetNet	10^{-3}	0.1543	no	-0.016	-0.032
	3×10^{-3}	0.9859	no	-0.016	-0.032
	10^{-2}	0.9062	no	-0.016	-0.032

GLA and Mamba-2 certify at every rate and RetNet at none (best 0.9859 at 3×10^{-3}), so its label stands; at 3×10^{-3} the second GPU replays the earlier accuracies within 0.001. The slowest channel does not depend on the rate — GLA's is zero at every rate, Mamba-2's -0.003 to -0.001 — and the fastest does: GLA's runs from -1.3 at 3×10^{-3} to -1.6 at 10^{-3} and -3.1 at 10^{-2}, and Mamba-2's span collapses to -0.0033 to -0.003 at 10^{-2}. By the rule declared before the sweep, Section 9 and Figure 3b keep the published rate, at which both certified baselines also score highest; the fastest-channel values are properties of a model at its rate, not of an architecture. At sequence length 512 Mamba-2 and RetNet stayed at chance and GLA reached 0.15 to 0.73, so the axis is reported on the $N = 8$ cell, where GLA, Mamba-2 and DeltaNet pass the ≥ 0.99 gate. DeltaNet's own $N = 48$ and $N = 64$ checkpoints were trained with that gate off (0.951 and 0.522) and are called certified only in the sense of the GPU/CPU agreement of Table 9. The isotropic-key reference is the analytic approximation $-\mathbb{E}[\beta(2-\beta)]/(2d_h)$, not a run with isotropic keys: $\|(I - \beta kk^\top)v\|^2 = 1 - \beta(2-\beta)\cos^2\theta$ with $\mathbb{E}\cos^2\theta = 1/d_h$, and the QR estimate is a rate on the norm, not on its square (`DELTANET_LYAPUNOV_20260920.json`).

DeltaNet at the decoding level. Per-item attribution fails at $N \geq 32$ because the read is not carried by one item. The decoding is still nearest-element: the rebuilt model's logits are $\langle z_t, h_w \rangle$ with h_w the output head's rows and z_t a per-query linear image of the memory read $o_t = S_t^\top q_t$ (RMS and layer normalisation are per-query scalars, taken at the actual read, so the map is exact for it). Split o_t into its projection onto the subspace spanned by the values written in the short-convolution window that starts at the true write — the window the scope test grouped by — and the rest, and apply Eq. 4 with the head rows as the frame and the residual's image as the crosstalk. On the four checkpoints, measured / predicted / union bound: $1.000/1.000/1.000$ at $N = 8$, $0.996/0.981/0.952$ at $N = 32$, $0.952/0.927/0.808$ at $N = 48$, $0.523/0.425/0.111$ at $N = 64$ — within 0.10 of measured, always from below, on a model whose own accuracy runs from 1.0 to 0.52. The window subspace holds 51 to 92% of the read's energy, falling with load, and the product-form coefficients say how it got there: at $N \leq 48$ the window's own writes contribute 4.1 to 7.7 times the projected read and the rest of the prefix cancels 3.3 to 6.8 times it, which is the delta rule's least-squares correction made visible and is why no single item dominates; at $N = 64$ the correction has collapsed (0.9 and -0.4) along with the accuracy. Two splits that fail are recorded: along the single stored value v_{s^*} the read has no component at $N \geq 32$ (the value is spread over the window by the convolution), and along the decoder's own direction the crosstalk is structured rather than Gaussian (Gaussian prediction 0.01 against 1.0 measured at $N = 8$). The relation therefore holds for DeltaNet at the decoding level, with descriptors that name no item; what it does not give is the factorisation into routing, gate and frame, which needs a stored item to attribute to.

The mixture-of-memories router. The official MoM layer (Du et al., 2026) with four memories, top-1 routing, no shared memory and no load-balancing loss (the package's defaults are eight memories, top-2, a shared memory and an auxiliary loss) was trained on the paper's four cells at vocabulary 512 (accuracy 0.600, 0.548, 0.458, 0.837; none certified, and not monotone in load) and

its router read with the paper's diagnostics. This is a diagnostic configuration with the collapse-prevention devices deliberately off, chosen so the routing statistics could be read on one memory table; it is not a claim about MoM as released, whose defaults were not run. The router is symbol-determined (every symbol goes to one memory on 100% of its occurrences) and collapsed: 86, 84, 80 and 94% of tokens go to one memory, against 52% at initialisation, with a mean top probability of 0.35–0.41. Among the stored keys of a sequence the collision rate equals the load-only value $\sum_m q_m^2$ to two decimals (0.681 against 0.679 at $N = 16$), and the soft mean square coherence sits *above* its per-symbol permutation (rank 200 of 200), which is what a collapsed load looks like under a permutation that spreads it. The diagnostic transfers; this router has nothing beyond load for it to read, and the four accuracies are those of a single memory doing the work.

Withheld learning on DeltaNet. Preregistered before any arm was trained (the registration ships with the artifact) and run once, on the certified recipe of each cell, seed 17: one parameter group frozen at its initialisation, the rest trained. Accuracy against the full model (the CPU rebuild of the certified checkpoint, Table 9):

frozen	N=8	N=32	N=48	N=64
nothing (full model, CPU rebuild)	0.9993	0.9957	0.9541	0.5247
keys and queries (addressing)	0.037	0.005	0.002	0.001
β (erasure gate)	0.9997	0.987	0.931	0.526
values (content)	0.994	0.429	0.271	0.081
output projection and head (decoding)	0.325	0.000	0.000	0.000

Addressing is learned at every load, and nothing compensates for it through the embedding; the erasure gate is not learned, in the sense that a random β costs 0.0, 0.8, 2.3 and 0.0 points; content matters from $N = 32$ on, and decoding absolutely from $N = 32$ on. This is the order of Figure 1a read against a diffuse initialisation — the routed memory's router is DeltaNet's addressing, and neither memory needs its gate learned as much as a diffuse addressing — and it is a parallel, not the same mechanism: the routed memory's cost falls to 0.1 to 1.4 against a fixed concentrated assignment (Appendix F), a comparison not run here; DeltaNet retrieves collectively, and its frozen keys are random directions in d_h dimensions that a single least-squares pass cannot separate. The relation of the previous paragraph, evaluated on these arms with the same window split (Figure 3a; measured accuracy, prediction, union bound and the crosstalk variance ratio, median over queries): Rows marked † hold cells past the tolerance of Appendix F: 95 of the 120 cells at $N = 8$ across the three metrics, none at $N \geq 32$; the per-cell flag is `_frac_below_tol` in the JSON, and no ratio below is built on a flagged cell.

frozen	N	measured	predicted	union bound	variance ratio
nothing (full model)	8	1.000	1.000	1.000	0.56
nothing (full model)	32	0.996	0.981	0.952	0.59
nothing (full model)	48	0.952	0.927	0.808	0.51
nothing (full model)	64	0.523	0.425	0.111	0.33
addressing	8	0.029	0.059	0.004	0.53
addressing	32	0.005	0.000	0.000	0.65
addressing	48	0.003	0.000	0.000	0.81
addressing	64	0.001	0.000	0.000	0.83
erasure gate β	8	1.000	0.999	0.999	0.56
erasure gate β	32	0.985	0.964	0.907	0.61
erasure gate β	48	0.930	0.884	0.722	0.47
erasure gate β	64	0.528	0.474	0.171	0.30
values	8	0.996	0.990	0.982	0.70
values	32	0.427	0.422	0.123	0.34
values	48	0.256	0.306	0.051	0.28
values	64	0.085	0.058	0.002	0.27
decoding	8	0.274	0.268	0.232	1.00
decoding	32	0.000	0.000	0.000	1.00
decoding	48	0.000	0.000	0.000	1.00
decoding	64	0.000	0.000	0.000	1.00

The prediction stays within 0.06 of the measurement on every arm and load; it is below the measurement on the full model at every load and on 14 of the 16 arm cells, and above it on addressing at $N = 8$ (by 0.03) and values at $N = 48$ (by 0.05). The variance ratio sits at 0.3 to 1.0 on the arms, as on the full model, so assumption (ii) is not what the two-sided error reflects; the arms whose accuracy is at chance are predicted at chance.

Condition numbers on DeltaNet. The fingerprint of Appendix F, read on DeltaNet with the read $o_t = \sum_{s \le t} \beta_s \rho_{s,t} v_s$ per head: geometry is the unit value direction as the model writes it ("no geometry" is the same seed's untrained floor), the gate is β_s, and the routing is the addressing $\rho_{s,t} = k_s^\top P_{s+1} \cdots P_t q_t$. The Gram of the prefix's values, $\sum_s a_s^2 \hat{v}_s \hat{v}_s^\top$ over its nonzero spectrum, is conditioned with the components switched on alone and in combination, medians over queries and heads, $\log_{10}$. Because DeltaNet fixes no norm on its values, three value metrics are read and all three are reported, none selected after the fact. No separate dated document declares them; the script that names all three was fixed on 2026-09-18, the day before the measurement was written, and ships in the artifact.

Euclidean: $\|v\|_2$, the values as the model writes them.

frozen	N	geom.	gate	routing	geom.+gate	geom.+routing	all	as read	floor
nothing (full model)	$8^\dagger$	7.7	6.7	9.7	8.5	11.8	11.2	12.2	7.3
nothing (full model)	32	2.1	2.2	3.5	2.1	3.9	4.0	4.0	2.5
nothing (full model)	48	2.2	2.0	3.1	2.2	3.7	3.7	3.4	2.4
nothing (full model)	64	1.8	2.0	2.4	1.9	2.2	2.4	2.5	2.3
addressing	$8^\dagger$	6.5	8.7	7.3	8.8	7.8	9.4	9.1	7.3
addressing	32	2.3	2.4	2.2	2.9	2.2	2.9	3.0	2.5
addressing	48	2.1	2.2	2.0	2.5	2.0	2.3	2.5	2.4
addressing	64	1.9	2.0	1.9	2.2	1.8	2.0	2.1	2.3
erasure gate β	$8^\dagger$	7.8	6.5	9.5	8.4	11.5	11.0	12.0	7.3
erasure gate β	32	2.2	2.3	3.3	2.4	3.9	4.0	4.0	2.5
erasure gate β	48	1.9	2.1	3.0	2.0	3.3	3.4	3.2	2.4
erasure gate β	64	1.9	2.0	2.4	2.1	2.4	2.5	2.6	2.3
values	$8^\dagger$	6.2	8.2	9.3	8.4	9.5	10.7	10.6	7.3
values	32	2.4	2.2	3.1	2.4	3.1	3.4	3.4	2.5
values	48	2.1	2.0	3.3	2.1	3.3	3.5	3.5	2.4
values	64	1.9	1.9	2.5	1.9	2.6	2.7	2.7	2.3
decoding	$8^\dagger$	9.7	7.5	10.0	11.1	12.8	13.5	14.0	7.3
decoding	32	2.2	2.4	2.3	2.6	2.3	2.7	2.8	2.5
decoding	48	2.1	2.2	2.0	2.4	2.1	2.3	2.4	2.4
decoding	64	2.0	2.0	1.9	2.2	1.9	2.1	2.2	2.3

decoder: $v^\top G v$ with G the head block of $W_o^\top P_c H^\top H P_c W_o$, the metric in which the head reads a value.

frozen	N	geom.	gate	routing	geom.+gate	geom.+routing	all	as read	floor
nothing (full model)	$8^\dagger$	8.9	8.8	11.2	9.5	12.7	12.2	13.1	9.4
nothing (full model)	32	4.3	4.2	5.2	4.3	5.3	5.3	5.6	4.4
nothing (full model)	48	3.9	4.0	5.3	3.8	5.0	5.0	5.0	4.7
nothing (full model)	64	2.7	3.0	3.7	2.7	3.4	3.6	4.1	3.5
addressing	$8^\dagger$	6.7	9.6	8.6	8.8	7.9	9.6	9.6	8.5
addressing	32	3.1	3.9	4.5	3.3	3.7	4.0	4.5	4.6
addressing	48	3.1	3.7	4.2	3.2	3.4	3.6	4.2	4.4
addressing	64	3.1	3.8	4.1	3.3	3.5	3.7	4.2	4.2
erasure gate β	$8^\dagger$	9.0	8.5	11.1	9.3	12.4	12.0	12.9	9.4
erasure gate β	32	3.9	3.9	4.7	3.9	5.0	5.1	5.5	4.0
erasure gate β	48	2.9	2.6	3.8	2.9	4.1	4.1	4.3	3.3
erasure gate β	64	2.8	2.7	3.3	2.9	3.3	3.4	3.7	3.2
values	$8^\dagger$	8.3	10.3	11.0	10.4	11.1	12.4	12.1	9.2
values	32	4.4	4.2	4.8	4.4	4.9	5.1	5.2	4.6
values	48	4.5	4.5	5.4	4.4	5.4	5.5	5.6	4.8
values	64	4.0	4.1	4.5	4.0	4.7	4.7	5.1	4.5
decoding	$8^\dagger$	10.2	7.8	10.3	11.7	13.3	13.9	14.5	7.7
decoding	32	3.1	2.8	2.7	3.4	3.1	3.5	3.6	2.9
decoding	48	2.9	2.6	2.5	3.1	2.8	3.0	3.2	2.8
decoding	64	2.7	2.5	2.3	2.9	2.7	2.8	2.9	2.7

whitened: $v^\top \Sigma^{-1} v$ with Σ the value covariance over the stream.

frozen	N	geom.	gate	routing	geom.+gate	geom.+routing	all	as read	floor
nothing (full model)	$8^\dagger$	8.4	7.5	10.5	9.1	12.1	11.7	12.3	8.0
nothing (full model)	32	2.2	2.4	3.7	2.2	4.1	4.1	4.0	2.7
nothing (full model)	48	1.9	2.2	3.3	1.9	3.6	3.6	3.4	2.6
nothing (full model)	64	1.6	2.2	2.6	1.8	2.4	2.4	2.5	2.6
addressing	$8^\dagger$	5.5	9.2	8.0	7.8	6.7	8.4	9.0	8.0
addressing	32	2.1	2.9	2.7	2.7	2.0	2.6	2.8	3.0
addressing	48	1.7	2.6	2.5	2.2	1.7	2.1	2.2	2.7
addressing	64	1.4	2.3	2.3	1.8	1.5	1.8	1.9	2.6
erasure gate β	$8^\dagger$	8.3	7.3	10.2	8.8	11.8	11.5	12.1	8.0
erasure gate β	32	2.3	2.4	3.4	2.5	4.0	4.1	4.0	2.7
erasure gate β	48	1.7	2.2	3.1	2.0	3.3	3.4	3.2	2.5
erasure gate β	64	1.7	2.2	2.6	2.0	2.4	2.5	2.6	2.5
values	$8^\dagger$	5.4	7.5	8.9	7.5	8.9	10.1	10.3	6.7
values	32	1.6	2.3	3.2	1.8	3.0	3.4	3.5	2.5
values	48	1.3	1.8	3.4	1.6	3.2	3.4	3.5	2.1
values	64	1.2	1.6	2.5	1.4	2.4	2.6	2.7	1.9
decoding	$8^\dagger$	9.2	8.0	10.4	10.6	12.6	13.2	14.1	7.8
decoding	32	1.6	3.1	3.0	2.3	1.8	2.5	2.6	3.2
decoding	48	1.3	2.9	2.8	2.2	1.5	2.2	2.4	3.1
decoding	64	1.2	2.8	2.8	2.0	1.4	2.1	2.2	3.1

Three readings hold under all three metrics. Addressing is what moves the conditioning on the full model. Over the three loads where κ resolves it moves it by $3.6\times$ to $22\times$ (Euclidean), $7.3\times$ to $29\times$ (decoder) and $11\times$ to $33\times$ (whitened), while the erasure gate moves it by at most $4.1\times$; the $N = 8$ values, which ran to $229\times$, are past resolution († rows) and are not used; that is the order of the withheld arms above, keys essential and β free. With addressing frozen the gate takes over at $N = 8$, in stable rank rather than in κ, which is past resolution at that load in every arm and metric. On DeltaNet the two statistics also disagree where both resolve — at $N = 64$ the Euclidean κ ranks the geometry above the gate and the stable rank ranks them the other way —

so we state the component order here less firmly than on the routed memory, where they agree in every row. At $N \geq 32$ the gate moves it by $1\times$ to $8\times$ and the routing by $1\times$ to $23\times$ (the larger values under the decoder metric). And the model as it reads is worst conditioned where it is most accurate. At $N = 8$ we say that in stable rank, the condition number being past resolution there: the read state carries 1.2 to 1.3 of a rank-64 prefix across the three metrics, against 2.5 to 5.9 under the trained geometry alone, so what the query decodes from is effectively one direction. At $N = 64$, where κ is a measurement, the as-read column is $10^{2.5}$ (Euclidean) and within $2\times$ of its floor. Two readings differ from the routed memory. The geometry column is not invariant across arms here, because DeltaNet's value directions are trained in every arm but one, whereas the routed memory's codebook directions barely move; at $N = 64$, where the statistic resolves, the span is 0.2 to 1.3 decades. And one reading is metric-dependent and therefore carries nothing: with decoding frozen, at $N \geq 32$ the gate multiplies the conditioning by $37\times$ to $44\times$ under the whitened metric and by $0.5\times$ to $1.4\times$ under the other two (at $N = 8$ it divides it under all three). Files: `KAPPA_COMPONENTS_DELTANET_20260918{,_decoder,_whitened}.json`.

Table 9: The pre-registered scope test on DeltaNet. The model is accurate and the instrument reads the wrong sign. The three rules are these. Write m_s for the magnitude of position s's contribution to the read, $W(z)$ for the union of convolution windows over every position holding token z, κ and w^* for the queried key and its answer, and $\mathcal{K}$ for the other stored keys of the sequence. With $g^* = \sum_{s \in W(\kappa) \cup W(w^*)} m_s$ and $g_k = \sum_{s \in W(k)} m_s$, the first rule is $C_{\text{all}} = g^{*2}/\overline{g_k^2}$ over $k \in \mathcal{K}$; the second replaces g^* by the mass of the single window opening at the last position holding w^*; the third is the signed margin $g^* - \max_{k \in \mathcal{K}} g_k$. Three independent attribution rules were tried, each justified before it was run; all three fail. The CPU rebuild reproduces the GPU-certified accuracy, so this is not an implementation artefact.

pairs N	accuracy, CPU rebuild (certified)	C (target window)	C (single)	margin negative on
8	0.9993 (0.9996)	4.44	6.91	84.0%
32	0.9957 (0.9950)	18.53	29.89	94.5%
48	0.9541 (0.9510)	35.01	44.12	91.8%
64	0.5247 (0.5216)	45.39	55.16	90.6%

The fingerprint on a released model trained on text. The fingerprint is the one reading in this paper that needs no training and no readout identity, so it can be taken on a public checkpoint. We take it at three depths, layers 0, 11 and 23 of `delta_net-1.3B-100B` from the official library (Yang & Zhang, 2024) — released at 1.3 billion parameters and trained on text, against this paper's 73,728 — on two unrelated text sources, under the Euclidean and whitened metrics (`KAPPA_COMPONENTS_FLA_DEPTH_20260919.json`, a later pass). Reading a layer other than the first means transcribing the block around the official recurrence, so that transcription carries a falsifier declared before it ran: token perplexity below 30 on every source. It reads 12.0 and 4.5.

Routing leaves the state a query decodes from more concentrated than the gate does in all 96 rows — three layers, two metrics, two sources, eight prefix lengths — on the stable rank, which is readable everywhere, and in every condition number that is a measurement. That ordering is what transfers, and it does not change with depth. Its size does: the two are 4.8 to 10.5 decades apart at layer 0 and 3.1 to 3.4 at layers 11 and 23, converging from both sides without crossing. So does the resolution, in the opposite direction — at layer 0 the value directions are nearly collinear and the Gram goes numerically singular once the prefix approaches the head dimension, which is where all 96 of the 768 unreadable condition numbers sit, while at the two deeper layers the residual stream has mixed and none does.

Two things do not, and both are scope rather than disagreement. The geometry column is not near its floor here: it reads 6.3 against a floor of 3.3 (Euclidean) and 3.4 against 6.8 (whitened), the opposite signs of one fact, that the trained value directions are strongly anisotropic and whitening by their own covariance removes exactly that anisotropy. This is not the contrast of Section 4, which is gauge-invariant and read against frozen arms; one released checkpoint has no arms, and the floor here is a draw from the released initialiser's own recipe rather than this checkpoint's own initialisation, so the two columns are not comparable and we do not compare them. The stable rank is the half that does not depend on the metric: the gate column carries 10.9 to 17.9 of 128 directions while every

trained-direction column sits at 1.5 to 1.8, so the state a query decodes from is effectively of rank one and a half once the trained geometry is in play.

What this buys is one sentence, and we do not extend it: the component ordering is readable at that scale on real text. It says nothing about accuracy or capacity there, and nothing about withheld learning, which needs arms this checkpoint does not have. Two limits are recorded with it. At prefixes near the head dimension the Gram is numerically singular and κ is pinned to the clamp rather than measured; the JSON marks every cell with whether κ is a measurement, and the slices quoted here are clean. And the decoder metric of Appendix G is not ported, because it centres as a layer normalisation does and this library normalises without centring, so porting it would be an invention rather than a reading.

Where the transition takes over. One quantity above is measured past this paper's reach on purpose. The stationary state has a conditioning of 325 to 2,020 against 6 to 79 for independent writes, and no query here decodes from it: every sequence is 512 tokens or fewer, and the initial state is forgotten over 1,500 to 6,300. Past that horizon the state a query reads is the stationary one rather than the transient one the decoding analysis uses, and its conditioning — measured, and not explained by correlated writes — is the object the fingerprint would then read. That is where this analysis ends.

H Record of corrections

This appendix exists because several numbers in an earlier draft of this work were wrong, and the corrections are part of what the reader is entitled to see. It is also the evidentiary basis for the disclosure statement on the use of large language models.

Wrong distractor population. An earlier routing contrast was measured against the memory's full 512-word vocabulary rather than the sequence's stored keys, giving ratios of 0.36–0.78 instead of 6.1–7.9 and inverting the sign of the conclusion. The measurement is retained in the artifact, marked superseded, because the contrast between the two populations is instructive.

An invalid power claim. A claimed separation of "18 standard errors" divided a log second-moment ratio by the standard error of a log contrast ratio. Those are different scales and the quotient is meaningless. Withdrawn, along with the null it was defending: the effect it claimed to bound is present at $|t| \geq 37$ once estimated exactly. The withdrawal was at first recorded only in the output file and not in the script that writes it, so re-running the measurement reinstated the retracted sentence; the invalid quotient is now deleted from the code.

A single draw of a heavy-tailed ratio. One reported contrast came from a single draw and read 1.63; 512 draws of the same quantity gave a median of 23.4 with standard deviation 256.4. All reported values are now medians with the draw count stated.

The wrong binding convention. An early operator instrument formed the term $\langle A(x_t)^\top A(k_i)B(v_i), B(w)\rangle$ — key i bound to value i — over key/value pairs drawn at random from the vocabulary. The verification of Eq. 1 showed that the member binds the key at $s-1$ to the value at s, and that the relevant distractor population is the sequence's stored keys rather than random draws. Under the corrected convention the statistic reads 2.1–2.6$\times$ the random-binding floor; under the old one it read 0.20–0.70$\times$, i.e. the opposite side of the floor. Numbers from the superseded instrument appear nowhere in this paper, and the episode is why Section 2 spends a paragraph on specifying the distractor population and the gauge before any contrast is quoted.

An omitted convolution. The DeltaNet extraction initially skipped the model's short convolution. It was caught because the resulting signed margin was negative on a model that answers correctly 99.96% of the time — that is, by the sanity condition rather than by inspection — and a reconstruction gate was added, which now reads exactly 0 against the official reference implementation.

A silent generator fallback. The harness contained a fallback from the official generator to a locally certified numerical port. The provenance guard that was supposed to prevent this had never run, because the record it reads was missing. Every routed-memory number in this paper is now measured on the official package (the DeltaNet arm and the baselines use the artifact's numerical port, stated where they appear); no conclusion changed. The readout-identity check was the last holdout and was re-run, moving its worst relative error from 1.0093×10^{-15} to 1.0683×10^{-15}.

The largest observed difference on any re-measured quantity was 0.008 in accuracy, and the scaling exponent moved from 1.023 to 1.026.

Results that lived only in their output files. Two quantities the paper quotes — the location of the temperature cliff and the scaling exponent — had been computed once and written by hand into the JSON artifacts rather than emitted by the scripts that produced them. Re-running on the official generator destroyed both. They are now derived in code from the measured rows, and re-deriving the cliff exposed an error in the hand-written version: it reported the cliff as $\tau : 1 \to 2$ for all four checkpoints, which is true only of the two lowest loads. At $N = 48$ and $N = 64$ the largest drop is one doubling later. A claim stated in prose beside a measurement, rather than computed from it, is a claim nothing checks.

An artifact that passed its own gate while leaking and while being unreproducible. The anonymisation gate for the supplementary matched its search terms with word boundaries. Pickle length-prefixes its strings, so the byte beside the internal name inside a checkpoint is a word character and the boundary could not match; four checkpoints carrying that name passed clean. The gate also never examined file *names*, and those checkpoints were named after the project. Both are fixed, and re-running the repaired gate on the package it had just certified returned 102 further findings, because the renaming step contained the identical boundary assumption in three of its rules. The same mistake occurred four times in this work before it was named. Independently, the packaging step had been dropping checkpoint files, which left an artifact that satisfied every check and could not reproduce a single number in the paper; checkpoints now ship, with every tensor hashed before and after rewriting.

A ceiling fitted to the exam it then marked. The first version of Section 4's reachable-range measurement optimised the codebooks on the same sequences it then read the floor, the trained model and the optimum on. It reported headroom rising to $1.76\times$ the floor and the claim that training had left a 76% operator gain unclaimed, and that claim reached a draft abstract. Optimising on disjoint sequences removes the effect entirely (0.99–$1.01\times$), and the in-sample-over-held-out ratio tracks the apparent gain almost exactly. Two checks were missing and are now present: the contrast reads only along $B(w^*)$, so an optimum is free to stop discriminating among the other 511 candidates, and it does — accuracy falls by 27.8 to 30.0 points from the trained codebooks and by 43.2 to 49.9 from the random ones. The claim is withdrawn.

And a replacement that was itself too strong. The withdrawal above was first written as "there was nothing to take": on this statistic the random floor was said to be effectively the ceiling. That was a property of optimising on a fixed set. Resampling the objective at every step lifts the contrast to 1.38–2.03 times the floor on sequences it has never seen, so there was something to take after all. What survives, and is stronger, is that taking it costs 21.7 to 26.6 accuracy points, and that a frame-potential bound at the measured same-key overlap leaves 5 to 11% of contrast reachable while the values sharing a key stay as separable as a random binding leaves them, so a gain of 1.4 to $2.0\times$ must raise that overlap. The bound does not show the cost is forced; an earlier wording said so and was withdrawn. The null is now bounded by an argument instead of by a search.

A relation whose first variance double-counted a term. The first version of the crosstalk model behind Section 2.4 wrote the interference on a competitor's *margin* against the difference vector $W^* - W(k, w)$, which carries the true write's own crosstalk $\langle N_t, W^* \rangle$ — a term common to every competitor and not a source of variance between them. The predicted variance came out 2.9 times the measured one, and the prediction undershot accuracy by up to 0.44. The check that caught it — measured over predicted crosstalk variance, reported per cell — is the one Table 7 now carries. The model was rewritten on the logits, with the common term integrated out, before any number from it entered this text.

A test set that was not the test set. The official generator draws its filler tokens from the global random-number generator, so a test set generated after a fresh seed differs from the one the training script scored after drawing its training set. The first relation run reproduced the recorded accuracies only to filler noise; the released script replays the training draw and reproduces the arm accuracies exactly and the four full checkpoints to within 0.0022 of their training-time readings (0.9699, 0.9363, 0.7998, 0.6931 against 0.9717, 0.9378, 0.8000, 0.6909).

A penalty term that carried no gradient. The accuracy-constrained search of Section 4 added λ times the task cross-entropy to its objective, and computed that cross-entropy by copying the

optimised codebooks into the frozen model with an assignment that does not differentiate. The term was therefore a constant: every λ arm descended pure contrast, and the spread across λ that looked like a trade-off was filler noise, the generator drawing its non-query tokens from the global random stream so that consecutive arms saw different fillers. Re-running with the penalty differentiable leaves the claim standing at a smaller number, $1.03\times$ rather than $1.07\times$ within two accuracy points. The check that would have caught it, that an added term's gradient is not identically zero, now runs before the first step of every arm.

An accuracy cost that was partly the step size. The same search's cost was read as the price of the objective. Descending the task loss alone at the same rates costs accuracy too, and at $N = 16$ costs more, so the attribution was wrong at that coordinate and partial at the others. Section 10 now says so.

A same-key overlap asserted, not measured. Section 4 said the resampling search took its gain in the cross writes, with the same-key overlap left at its random value, while Appendix D said a gain of that size had to raise it. Neither was measured at the optimum. It is now: the overlap rises to 1.78 to 1.85 times the trained binding's and above the random binding's on all four checkpoints, so the body sentence is withdrawn; the appendix's inference that the gain was bought with that rise is withdrawn in its own entry below. Each was half right.

An accuracy cost read from one permutation. Earlier text gave the accuracy cost of destroying arrangement as one per-symbol permutation per checkpoint (seed 18): 0.0002, 0.0236, 0.0324, 0.0409. Over 200 permutations the means are 0.005, 0.009, 0.023, 0.031, and the $N = 24$ draw lay outside the range of all 200. The text now quotes the means, the ranges, and the cost against load-matched nulls, which is smaller again (Table 5).

An over-correction. One correct result was marked void on the grounds that its operator term did not match the notation of an earlier appendix. The verification of Eq. 1 later showed the term was right and the notation was wrong. The voiding was retracted.

A bound quoted outside its regime. Appendix D said a ten per cent contrast gain at $K = 64$ needs $c_s \geq 0.1016$, three times the random level. The threshold was evaluated at the isotropic $c_d^{\mathrm{kv}} = c_s = 1/d$ while the paragraph's inputs are the measured coherences, at which it is 0.069, 2.2 times; and the search optimum raises c_d^{kv} rather than lowering it, so the inequality does not relate the gain to the same-key overlap. Rerun with the optimum saved, the contrast read on stored pairs alone is 0.81 to 0.90 of the floor: the gain sits in the artefact writes, which Section 4 and the appendix now say.

An isotropic reference in the wrong convention. The Lyapunov exponents are QR estimates of the rate on a vector's norm; the isotropic-key reference $-\mathbb{E}[\beta(2-\beta)]/d_h$ is the rate on its square. The reference is now $-\mathbb{E}[\beta(2-\beta)]/(2d_h)$ with both values kept in the artifact, and "twelve to sixty-five times slower" is six to thirty-two.

Three sentences sharper than their evidence. Appendix F said only 3 of 36 penalised arms improve on their $\lambda = 0$ arm; 33 recover accuracy and 3 raise the contrast, and the sentence now counts both. Appendix D used the relation's symbol c_d for the stored-pair coherence its bound constrains, which is a different population; it is now c_d^{kv}. Seven routed and ninety-five DeltaNet condition numbers past the tolerance were printed unmarked beside the sentence that says none are quoted; they are marked, and the paragraph says which quantity the $1/\sqrt{\epsilon}$ line applies to.

An operator ratio read at its own resolution. The cosine-gauge ratio was 1.010 to 1.018 over 512 paired draws, whose two-standard-error bound was 1.037 to 1.078: the reading sat inside its own noise. At 4096 draws, under a rule declared first, it is 0.997 to 1.002 with a bound of 1.014 to 1.021, now inside the room the frame bound leaves. The null is unchanged and its statement is now bounded.

A routing cost read against the weakest fixed routing. The frozen router of Section 3 was the initialiser's draw, and the 37 to 55 points it costs were read as what learned routing is worth. Against a frozen random hard hash the cost is 0.1 to 1.4 on three seeds. The text now carries both, leads with the hash, and says what a concentrated, balanced assignment supplies.

I TERMS USED THROUGHOUT

Table 10: Terms used throughout. The first four rows are the statistics of Section 2.2.

Term	Meaning
$\mathrm{Orth}(A)$	$\|A^\top A - I\|_F / \|I\|_F$, the customary orthogonality deviation; not gauge invariant (Section 2.3).
C_{op}	$m_{s*}^2 / \mathbb{E}_s[m_s^2]$: the operator contrast of the target against the distractor population; read raw and in the cosine gauge.
C_{route}	$g_{s*}^2 / \mathbb{E}_s[g_s^2]$: the routing contrast.
C_{full}	$c_{s*}^2 / \mathbb{E}_s[c_s^2]$ on $c_s = m_s g_s$: split-invariant; the primary quantity.
floor	The same architecture, configuration and seed at step 0, evaluated on identical draws, every learned tensor at its initialisation.
write types	The shift convention produces *true pairs* (key at $s-1$, value at s), *cross writes* (a value bound to the next key) and *query-side* writes.
write gate	The per-write norm by which training separates true pairs from the other two types, read from token identity (Section 4).
concentration	How few slots a write lands on: effective slot count $\exp H$, participation ratio $1/\sum_i g_i^2$.
arrangement	Which symbols share slots at fixed per-symbol concentration: mean square overlap $\mathbb{E}[b^2]$ of pairwise gate overlaps b, read against per-symbol slot permutations (Section 6).
load	Frequency-weighted slot usage $\|\sum_i p_i g_i\|^2$, with uniform minimum $1/M$.
hash router	A random balanced one-hot assignment, 32 symbols per slot, frozen.
diffuse router	The initialiser's frozen draw, about 10.5 effective slots.
descriptors	Per-query routing weights, write norms and same-key overlaps ρ_w: the inputs of Eq. 4.