THE KV CACHE AS A ROTARY ASSOCIATIVE MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

KV cache compression is standard practice in long-context inference. Existing compression methods are validated by benchmarking after the fact, providing no a-priori guarantee on retrieval performance. Across ten models and four compression methods, retrieval accuracy falls by 41% on average, while zero-shot accuracy falls by 2% and perplexity rises by 19%. To understand the structure responsible, we formulate every attention read from the KV cache under rotary position embeddings (RoPE) as an exact retrieval from an associative memory, whose stored addresses are related across read times by a unitary representation of time translation at a fixed spectrum of frequencies. We term this view a **rotary associative memory** (RoAM). The representation decomposes the cache into frequency bands and gives closed forms for how far a head can retrieve (retrieval range), how many entries it separates at that distance (capacity at distance), and where distinct positions become indistinguishable (positional aliasing). Using band-targeted perturbation, we show that retrieval at distance is carried by the slow bands. Removing an eighth of the slow bands raises perplexity to between 260 and 5.4×10^6 and drives retrieval to zero across RoPE bases from 10^4 to 10^7, whereas removing an eighth of the fast bands leaves retrieval within 0.045 of baseline. At matched perturbation norm, slow-band damage costs more retrieval accuracy on every model, while zero-shot accuracy moves by at most 0.010. Position-uniformity of the readout bounds the worst-case effect of any key edit at every future read position, with the edit-side factor computable at compression time and the query-side factor bounded from the model's weights. The bound is audited on 115 million measured perturbations from deployed reconstruction methods. Predicted attention-logit curves match measured ones head by head across the ten evaluated models, with coefficients estimated from activations and no parameters fitted to the curves (median detrended $r = 0.894$), and aliasing distances predicted in advance from each configuration alone are reproduced to the integer.

1 INTRODUCTION

In autoregressive inference, every new token is conditioned on the key-value (KV) cache, which stores all past key and value projections, serving as the model's only record of context. To reduce its footprint, the cache is routinely compressed: quantization (Hooper et al., 2024; Liu et al., 2024), eviction (Zhang et al., 2023; Liu et al., 2023), and low-rank projection (Chang et al., 2024), each perturbing the stored record differently. Existing methods are validated by benchmarking after the fact, on metrics that do not quantify the damage: across ten models and four compression methods, retrieval accuracy falls by 41% on average from a 0.592 identity baseline, while zero-shot accuracy falls by 2% and perplexity rises by 19%. Retrieval degrades more than zero-shot accuracy in 39 of the forty (model, method) pairs we evaluate, and in none does the reverse hold. So, the metrics reported alongside these methods understate what they cost.

To identify what retrieval depends on, we start from the rotation that writes position into the cache: under rotary position embeddings (RoPE), a key k_s at position s is stored as $\rho(s)\, k_s$ and a query q_t is presented as $\rho(t)\, q_t$ (Su et al., 2024), so the score depends only on content and on the offset $t - s$, with that dependence governed entirely by the spectrum. Barbero et al. (2024) characterize what makes rotary encoding useful, and Xu et al. (2024) bound usable context by the RoPE base. Neither analyzes the per-head quantities that decide retrieval: how far a head can retrieve (retrieval range), how many entries it separates at that distance (capacity at distance), and where distinct positions

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.

become indistinguishable (positional aliasing). Without them, the cost of a compression method cannot be bounded before it is deployed, and its effect on retrieval cannot be anticipated.

KV cache retrieval is one step of the concave-convex update on the energy of an associative memory (Hopfield, 1982; Ramsauer et al., 2020), which carries no position. We show that the missing structure is a group action: each key written at time s has an address, realized in the cache as $\rho(s)\, k_s$, and every read relates to the stored addresses through a unitary representation of time translation at the RoPE frequency spectrum. We call this view a **rotary associative memory** (RoAM), in which addresses are compared but never resolved, so no positional metadata is stored (Figure 1).

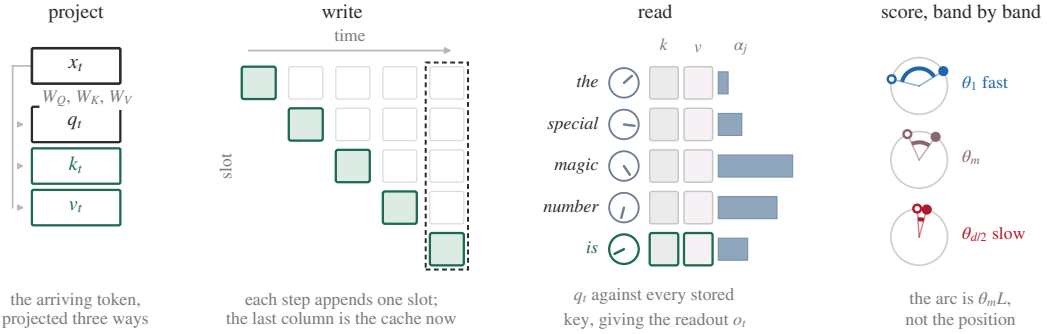


Figure 1: One decoding step of the KV cache as a rotary associative memory: the arriving token is projected to q_t, k_t, v_t; each step appends exactly one slot at address $\rho(t)$ (green), the boxed column being the cache as it stands; q_t is scored against every stored key for the weights α_j and the readout $o_t = \sum_j \alpha_j v_j$; and each score decomposes band by band, where only the arc $\theta_m L$ enters.

Under RoAM, each frequency θ_m in the spectrum contributes a term in $\cos(\theta_m L)$ to the attention score at distance L. Frequencies with $\theta_m L \gg 1$ oscillate through many periods and cancel out, so only the slow ones remain coherent at long range. Therefore, any structure a head applies at distance lives in the slow frequencies. Our **contributions** are as follows.

- A causal localization of retrieval at distance to the slow frequencies. Damaging them destroys retrieval, whereas damaging the fast frequencies does not, at a matched perturbation norm, so placement rather than magnitude accounts for it.
- A perturbation certificate bounding the effect of any key edit on attention. The bound is position-uniform, so one computation at compression time covers every future read position and every query; the edit-side factor is known then, the query-side factor is bounded by the weights, and a margin condition turns it into a guarantee on the retrieved entry.
- A capacity law for how far back a head retrieves and how many entries it separates there, expressed through a *signal kernel*, a closed-form function of the spectrum and the head's measured energy profile, together with offsets at which a spectrum renders distinct positions indistinguishable, computable from the spectrum alone and certified by finite search.
- A proof that the frequency schedule deployed in RoPE models is the unique spectrum meeting a scale-free coverage floor with maximal resolution: coverage across log-distance is a budgeted resource, and no other schedule dominates it. An equivariance theorem restricts position-consistent linear cache edits to band-diagonal filters.

We evaluate across ten models spanning six RoPE bases and three head dimensions. Predicted attention-logit curves match measured ones head by head, with no parameters fitted to the curves. Aliasing distances predicted from configuration alone are reproduced to the integer. The certificate holds without violation across the three deployed reconstruction methods we audit. Caching in RoAM form leaves perplexity and downstream accuracy unchanged within measurement noise.

2 The Rotary Associative Memory (RoAM)

2.1 Setup and Notation

Let d be even. Identify $\mathbb{R}^d \cong \mathbb{C}^{d/2}$ via $z_m = v_{2m-1} + i\, v_{2m}$, so that $\langle u, v\rangle_{\mathbb{R}} = \mathrm{Re} \sum_m \bar{u}_m v_m$. Fix a spectrum $\Theta = (\theta_1, \ldots, \theta_{d/2})$ of distinct frequencies $\theta_m \in (0, 1]$. The *geometric comb* with base b is the spectrum $\theta_m = b^{-2(m-1)/d}$, the frequency schedule of deployed RoPE models.

Representation. Define $\rho\colon (\mathbb{R},+) \to U(1)^{d/2}$ by $\rho(\Delta)\, z = (e^{i\theta_m \Delta}\, z_m)_m$. This is a unitary representation of the time-translation group: $\rho(0) = I$, $\rho(s)\,\rho(t) = \rho(s+t)$, and $\rho(t)^{-1} = \rho(-t)$. Its irreducible components are the one-dimensional $e^{i\theta_m \Delta}$, and the m-th coordinate is the m-th *band*.

Memory and readout. A *rotary associative memory* (RoAM) is an ordered tuple $C = (k_1, \ldots, k_t)$ of pre-action keys $k_j \in \mathbb{C}^{d/2}$. Position is carried entirely by the address (the write index); no metadata is stored. For a query q at time t, the pre-softmax score of entry j is

$$s_j(q,t) \;=\; \langle \rho(t)\, q,\ \rho(j)\, k_j \rangle_{\mathbb{R}} \;=\; \mathrm{Re} \sum_m \bar{q}_m\, k_{j,m}\, e^{-i\theta_m L_j}, \qquad L_j := t - j \geq 0, \tag{1}$$

followed by $\mathrm{softmax}(\beta\, s)$ against the stored values. L_j is the *retrieval distance* of entry j.

Derived spectral objects. For a fixed pair (q,k), write $c_m := \bar{q}_m\, k_m = a_m\, e^{i\varphi_m}$ with $a_m \geq 0$. The score as a function of distance, $s(L) = \sum_m a_m \cos(\theta_m L - \varphi_m)$, is the pair's *tuning curve*. For a band-energy profile $(\sigma_m^2)_{m=1}^{d/2}$ with total energy $E_{\mathrm{tot}} = \sum_m \sigma_m^2$, the *signal kernel* $S(L) := 2\sum_m \sigma_m^2 \cos(\theta_m L)$ is the expected score of a content-matched pair at distance L (Theorem 3). The *coherent energy* $E_{\mathrm{coh}}(L) := \sum_{m:\, \theta_m L \leq 1} \sigma_m^2$ is the profile-weighted energy still constructively aligned at distance L, and the *coverage* $D(\Theta; L) := \#\{m : \theta_m \leq 1/L\}$ counts the coherent bands.

2.2 Operational Equivalence and Positional Structure

Lemma 1 (Operational equivalence)**.** *For all q, k, t, j: $\langle \rho(t)\, q,\ \rho(j)\, k \rangle = \langle q,\ \rho(j-t)\, k \rangle$. The RoAM readout and standard post-RoPE attention therefore produce identical scores.*

We work in pre-action coordinates throughout, where all positional information is carried by the rotation ρ. Perturbing this rotation has a controlled effect on the readout.

Proposition 1 (Positional modulus of continuity)**.** *For all $t_1, t_2 \in \mathbb{R}$: (i) $\|\rho(t_1) - \rho(t_2)\|_{\mathrm{op}} = \max_m 2\,|\sin(\theta_m(t_1 - t_2)/2)|$; (ii) for every pair (q,k), $|s(q,k,t_1) - s(q,k,t_2)| \leq \sum_m 2\, a_m |\sin(\theta_m(t_1 - t_2)/2)|$, with equality for phase-aligned pairs; (iii) if every logit is perturbed by at most ε, the softmax weights shift by at most $\beta\varepsilon/2$ in total variation.*

Position can be perturbed at bounded cost. It cannot be removed.

Theorem 1 (Conservation of position)**.** *Suppose a memory stores only the multiset $\{k_j\}$ and scores each stored element by its value and (q,t) alone. If duplicates $k_{j_1} = k_{j_2} = k$ occur at positions with $\theta_m(j_1 - j_2) \notin 2\pi\mathbb{Z}$ for some band m with $k_m \neq 0$, then no such readout reproduces rotary attention.*

Position must therefore reside in the stored state, in explicit metadata, or in the address; the rotary associative memory places it in the address. Theorem 2 characterizes the operations that preserve it.

Theorem 2 (Equivariant cache operations)**.** *For distinct $\theta_m \in (0,1]$ and a real-linear map M, the following are equivalent: (i) M commutes with the single generator $\rho(1)$; (ii) $M\rho(\Delta) = \rho(\Delta)M$ for all $\Delta \in \mathbb{Z}$; (iii) $M\rho(\Delta) = \rho(\Delta)M$ for all $\Delta \in \mathbb{R}$; (iv) $(Mz)_m = \lambda_m z_m$ for some $\lambda_m \in \mathbb{C}$.*

The equivariant edits are precisely the band-diagonal gain-and-phase filters. Cross-band mixing breaks equivariance, so a linear edit meant to survive re-positioning must act band by band, and a single integer shift already enforces this. Nonlinear edits such as quantization fall outside the class, and Proposition 1 bounds them instead. Section 6 combines the equivariance constraint and the perturbation bound into certified compression bounds. Proofs are presented in Appendix A.

3 Capacity at a Distance

The coherent energy $E_{\mathrm{coh}}(L)$ bounds the signal kernel deterministically.

Lemma 2 (Coherent-energy bounds)**.** *For any band-energy profile and all $L \geq 1$, $2\cos(1)\, E_{\mathrm{coh}}(L) - 2(E_{\mathrm{tot}} - E_{\mathrm{coh}}(L)) \leq S(L) \leq 2E_{\mathrm{tot}}$.*

Energy in low-frequency bands keeps $E_{\mathrm{coh}}(L)$ near E_{tot} out to large distances, producing long-range heads. Energy in high-frequency bands produces local heads.

Let $k_j = \Sigma^{1/2} g_j$, where $g_1, \ldots, g_N$ are i.i.d. with independent, mean-zero, unit-variance, K-sub-Gaussian entries and $\mathrm{Tr}\,\Sigma = 1$; offsets are fixed before the draw, the target k_J sits at distance L, and the query is the content echo $q = k_J$. Write $D_L = \bigoplus_m \cos(\theta_m L)\, I_2$, so that $S(L) = \mathrm{Tr}(D_L \Sigma)$, and set $V_{\mathrm{sig}} := \|\Sigma^{1/2} D_L \Sigma^{1/2}\|_F^2$, $v_* := \|\Sigma^{1/2} D_L \Sigma^{1/2}\|_{\mathrm{op}}$, and $V_{\mathrm{int}} := \sup_\Delta \mathrm{Tr}(\rho(\Delta)^\top \Sigma\, \rho(\Delta)\, \Sigma)$.

Theorem 3 (Capacity at a distance)**.** *(i)* Success: *if* $S(L) \geq C\big(K^2\sqrt{V_{\mathrm{sig}} \log(1/\delta)} + K\sqrt{\|\Sigma\|_{\mathrm{op}} \log(N/\delta)} + K^2 v_* \log(1/\delta) + \beta^{-1}\log(N/\delta)\big)$, *then β-softmax places mass at least* $1 - \delta$

on the target with probability at least $1-\delta$. *(ii)* Failure: *for Gaussian keys, band-isotropic* Σ *with* $(\operatorname{Tr}\Sigma^2)^2/\operatorname{Tr}\Sigma^4 \geq 16\log(2/\delta)$, $N \geq 9$, *and* $S(L) \leq \frac{1}{2}\sqrt{V_{\mathrm{int}}\log(N-1)} - u_A(\delta/2)$, *with* u_A *the signal-concentration width, the top-scoring entry is a distractor with probability at least* $1-\delta-e^{-\Omega(N^{3/4}/\sqrt{\log N})}$.

Retrieval succeeds when $S(L)$ clears the interference floor and fails below it, at a crossover kernel value of $\Theta(\sqrt{V_{\mathrm{int}}\log N})$ for near-flat band-isotropic profiles. The theorem describes the ensemble above; Section 7 tests its constants on Gaussian stores and on captured keys.

Geometric comb. For the flat profile $\sigma_m^2 \equiv 1/d$ (so $E_{\mathrm{tot}} = 1/2$), the signal kernel tracks $1-\log_b L$ over $1 \leq L \leq b$, giving the working capacity law

$$\log N_{\max}(L) \asymp D(\Theta; L)^2/d, \qquad D(\Theta; L) = \tfrac{d}{2}\left(1-\log_b L\right) + O(1), \tag{2}$$

valid in the temperature regime where the softmax term of Theorem 3 is not binding.

Corollary 1 (Base bound)**.** *Under the flat geometric comb, retrieval at distance* L *from* N *stored keys requires* $\log_b L \leq 1 - \Theta(\sqrt{\log N/d})$*: a lower bound on the base* b *given a target context length. It recovers the base-context bound of Xu et al. (2024) in direction but not in constant, from memory capacity rather than single-pair discrimination.*

Corollary 2 (Per-head prediction)**.** *For a trained head with measured per-band cross-covariance* $\hat{C}_m$ *(the empirical mean of* $\bar{q}_m k_m$*), the expected attention logit at distance* L *is* $\sum_m \left[\operatorname{Re}\hat{C}_m \cos(\theta_m L) + \operatorname{Im}\hat{C}_m \sin(\theta_m L)\right]$*, with coefficients that are second moments of the activation distribution and none fitted to the curves. The content-matched case* $\hat{C}_m = 2\hat{\sigma}_m^2$ *recovers* $S(L)$. Proofs are presented in Appendix B.

4 Aliasing and Resolution

Theorem 4 (Aliasing)**.** *(i) For every* $\varepsilon \in (0, 2]$*, there exists an integer* $\Delta^* \in [1, \lceil 2\pi/\varepsilon \rceil^{d/2}]$ *with* $\|\rho(\Delta^*) - I\|_{\mathrm{op}} \leq \varepsilon$*; consequently every pair* (q, k) *is* ε*-confused simultaneously:* $|s(q, k, L+\Delta^*) - s(q, k, L)| \leq \varepsilon\|q\|\|k\|$ *for all* L. *(ii) If a head's energy is supported on* r *bands, an* ε*-alias of the restricted action exists at some* $\Delta_r^* \leq \lceil 2\pi/\varepsilon \rceil^r$.

At full dimension the bound in (i) exceeds any practical context length, so full-spectrum aliasing is an existence statement, not a practical concern. The consequential case is (ii): band-sparsity $r \ll d/2$ brings positional confusion exponentially closer, so measured per-head sparsity determines where it becomes realistic. For a concrete spectrum, absence of any ε-alias below a given range is finitely certifiable by lattice search; ε is a fraction of the logit scale, since an ε-alias moves any score by at most $\varepsilon\|q\|\|k\|$. Section 7 reports the certified alias-free ranges of deployed bases.

Proposition 2 (Range–resolution trade-off)**.** *Consider a tuning curve with amplitudes* a_m*, independent uniform phases, and a readout* $y = s(L) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \tau^2)$. *(i) The offset Fisher information is* $I(L) = s'(L)^2/\tau^2$, *and* $\mathbb{E}_\varphi I = \frac{1}{2\tau^2}\sum_m a_m^2\theta_m^2$*, independent of* L*: localization precision is carried by high-frequency energy. (ii) With* $E_a := \sum_m a_m^2$ *and* $E_{\mathrm{coh}}^a(L) := \sum_{m:\,\theta_m \leq 1/L} a_m^2$*, for every* $L \geq 1$, $2\tau^2\,\mathbb{E}_\varphi I + (1-L^{-2})\,E_{\mathrm{coh}}^a(L) \leq E_a$*, with equality iff the spectrum is supported on* $\{1/L, 1\}$.

Coherent range (Theorem 3) resides in low-frequency energy. Alias-free range (Theorem 4) shrinks with the high bands that localize. Proposition 2(ii) quantifies the competition on an attainable frontier: no head maximizes retrieval range, positional resolution, and alias-free range at once, yet a single spectrum must serve all three. Section 5 identifies the unique spectrum meeting a scale-free coverage floor with maximal resolution. Proofs are presented in Appendix C.

5 Optimal Spectra

The spectrum is a design choice. By Lemma 2, retrieval at distance L depends on the energy in bands with $\theta_m L \leq 1$, and $D(\Theta; L)$ counts them. A spectrum covering few bands at L cannot retrieve there under any energy allocation. No distance scale is privileged, so the criterion must be invariant under rescaling the unit of distance: the coverage lost when distance grows by a factor c may not depend on the starting distance, which forces the profile to be affine in $\ln L$. Fix a design range $L_{\max} > 1$ and the scale-free coverage profile $g(L) := \frac{d}{2}\left(1 - \frac{\ln L}{\ln L_{\max}}\right)$. A spectrum Θ *dominates* g if $D(\Theta; L) \geq \lfloor g(L) \rfloor$ for all $L \in [1, L_{\max}]$. The profile g is not chosen to match the comb; that the geometric comb with base $L_{\max}$ realizes it is a conclusion of Theorem 5, which is a converse:

meeting this floor forces every order statistic to lie at or below the comb's, and no dominating spectrum retains more resolution. The integer part is forced: D is integer-valued, and real-valued domination is achievable only by spectra whose every frequency is at most $L_{\max}^{-2/d}$.

Theorem 5 (Coverage–resolution characterization)**.** *(i)* Budget: $\int_1^{L_{\max}} D(\Theta; L)\, d\ln L = \sum_m \min\big(\ln \frac{1}{\theta_m}, \ln L_{\max}\big) \le \frac{d}{2} \ln L_{\max}$*: coverage on the log-distance axis is a conserved resource. (ii)* Characterization: Θ *dominates* g *if and only if its order statistics satisfy* $\theta_{(k)} \le L_{\max}^{-(1-2k/d)}$ *for every* $k = 1, \dots, d/2$*. (iii)* Optimality: *the geometric comb with base* $L_{\max}$*, given by* $\theta_{(k)} = L_{\max}^{-(1-2k/d)}$*, is the unique coordinatewise-maximal dominating spectrum, and it attains the minimal budget expenditure among them:* $\int_1^{L_{\max}} D\, d\ln L = \int_1^{L_{\max}} \lfloor g \rfloor\, d\ln L = \frac{d-2}{4} \ln L_{\max}$*, tight against the coverage floor. It therefore maximizes every permutation-invariant functional increasing in the order statistics, including the flat-profile resolution of Proposition 2.*

The criterion is coverage, a property of the spectrum alone.

Corollary 3 (Certified capacity of the design)**.** *Any* g*-dominating spectrum with flat profile* $\sigma_m^2 \equiv 1/d$ *has* $E_{\mathrm{coh}}(L) \ge \lfloor g(L) \rfloor / d$*, so Lemma 2 certifies the signal floor* $S(L) \ge \frac{2(1+\cos 1)}{d} \lfloor g(L) \rfloor - 1$ *for all* $L \le L_{\max}$*, and hence the success guarantee of Theorem 3 wherever the floor clears its threshold, which requires* $\lfloor g(L) \rfloor \ge 1$ *and so terminates at* $L_{\max}^{1-2/d}$*. Among all spectra providing this certified guarantee, the geometric comb with base* $b = L_{\max}$ *uniquely maximizes positional resolution. The frequency schedule deployed in RoPE models is thus the unique spectrum that meets the scale-free coverage floor for the range set by its own base* b *with maximal resolution.* Proofs are in Appendix D.

6 Structural Guarantees and Robustness

This section turns Theorem 2 and Proposition 1 into deployment guarantees for compression.

Proposition 3 (Perturbation certificate)**.** *Let the stored key of entry* j *be edited in place to* $k_j + \delta_j$*, the cache index set unchanged (quantization, low-rank projection, linear filtering). For every query* q *at every future position, the entry's logit shifts by at most* $\|q\|\, \|\delta_j\|$*, independent of retrieval distance; if* $\|q\|\, \|\delta_j\| \le \varepsilon$ *for every edited entry, the softmax weights shift by at most* $\beta\varepsilon/2$ *in total variation.*

The edit-side factor $\|\delta_j\|$ is computable at compression time and uniform over future read positions; the query-side factor $\|q\|$ is bounded by the weights (Corollary 4), and a margin condition turns the logit bound into a guarantee on the retrieved entry (Corollary 5). The certificate applies to every in-place edit; only edits meant to survive re-indexing must be band-diagonal (Theorem 2). Eviction changes the cache cardinality, so they alter the softmax denominator rather than perturb a logit, and are classified rather than certified (Section 7).

Corollary 4 (Deployment-time query bound)**.** *For a pre-normalized block with* $q_t = W_q^{(h)} \mathrm{RMSNorm}(x_t) + b_q^{(h)}$ *and gain* γ*, every query of head* h *satisfies* $\|q_t\| \le \|W_q^{(h)}\|_{\mathrm{op}} \sqrt{d_{\mathrm{model}}}\, \|\gamma\|_\infty + \|b_q^{(h)}\|$*, computable from the checkpoint alone. Substituting it for* $\|q\|$ *in Proposition 3 gives a certificate with no dependence on future inputs.*

Corollary 5 (Retrieval preservation)**.** *Fix a query* q *at read time* t*, let* J *be its top-scoring entry, and let* $\mu := s_J - \max_{j \ne J} s_j > 0$ *be its margin. If every stored key is edited with* $\|q\|\, \|\delta_j\| \le \varepsilon$ *and* $\varepsilon < \mu/2$*, then* J *remains the top-scoring entry after the edit.*

Corollary 6 (Certified position quantization)**.** *Reading an entry at an offset within* $\pm w$ *of its true offset perturbs its logit by at most* $\sum_m 2a_m \sin(\frac{1}{2} \min(\theta_m w, \pi))$*, computable from a live cache.*

Compression perturbs stored content. The analysis also assumes keys carry no position, a premise trained models need not satisfy. Fix an orthogonal projector P_{pos} onto a position-carrying subspace and split $q = q^{\perp} + q^{\mathrm{pos}}$, $k = k^{\perp} + k^{\mathrm{pos}}$, with leakage fractions $\eta_q := \|q^{\mathrm{pos}}\|/\|q\|$ and $\eta_k := \|k^{\mathrm{pos}}\|/\|k\|$.

Theorem 6 (Leakage robustness)**.** *For every distance* L*:* $|s(q, k, L) - s(q^{\perp}, k^{\perp}, L)| \le (\eta_q + \eta_k)\, \|q\|\, \|k\|$*. The softmax weights shift by* $O(\beta\eta)$ *via Proposition 1, and the thresholds of Theorem 3 hold with* $S(L)$ *replaced by* $S(L) - C\eta$ *for success and* $S(L) + C\eta$ *for failure: leakage degrades capacity linearly in the measured fraction.*

Corollary 7 (Band-resolved leakage)**.** *Suppose* P_{pos} *is band-aligned, a selector of a band set* $\mathcal{M}_{\mathrm{pos}}$*; these are the equivariant orthogonal projectors (Appendix A). Only the logit contributions of those bands are then perturbed. If* $\mathcal{M}_{\mathrm{pos}}$ *is disjoint from the coherent set* $\{m : \theta_m L \le 1\}$*, the coherent-band contribution to the score at distance* L *is unchanged. Band-localized leakage is therefore benign for retrieval at the corresponding distances, and equivariant operations isolate it.*

Proofs are presented in Appendix E.

7 EXPERIMENTS

Models and capture. We evaluate ten base checkpoints, 1.7B to 8B parameters, spanning six RoPE bases (10^4 to 10^7), three head dimensions (64, 128, 256), and declared contexts from 4096 to 131072: Llama-3.1-8B (the one scaled spectrum) and Llama-3-8B (Grattafiori et al., 2024), Mistral-7B v0.1 and Mistral-7B v0.3 (Jiang et al., 2023), Qwen2-7B (Yang et al., 2024) and Qwen2.5-7B (Yang et al., 2025), SmolLM2 (Allal et al., 2025), Yi-6B (Young et al., 2024), Gemma-7B (Team et al., 2024), and Granite-3.1 (Granite Team, 2024). Keys and queries are captured pre-rotation on ten PG-19 test documents (Rae et al., 2020) of 8192 tokens each (4096 for Yi-6B), in bf16; each spectrum is read directly from its rotary buffer. All spectrum-determined quantities (coverage at context, kernel values, candidate alias integers) are functions of configuration alone.

Deployed spectra and the design map. Table 1 places each model on the design map: coverage at the declared context L_{ctx} ranges from 0 to 29 bands. At the extremes, Mistral-7B v0.1's spectrum retains zero coherent bands at its declared $L_{\mathrm{ctx}} = 32768$, though its 4096-token attention window bounds its retrieval independently of the spectrum, Qwen2-7B and Qwen2.5-7B hold $S(L_{\mathrm{ctx}}) = 0.002$ at 131072, and Yi-6B keeps 29 bands for a 4096 context. The scaled and unscaled pair Llama-3.1-8B and Llama-3-8B isolates the design lever: scaling holds coverage nearly constant (20 to 16) as L_{ctx} grows sixteenfold from Llama-3-8B to Llama-3.1-8B.

Table 1: Model roster and spectrum-determined quantities. b is the RoPE base, d the head dimension, and D and $S(L_{\mathrm{ctx}})$ the coverage and flat-profile signal kernel at each model's declared context L_{ctx}.

Model	b	d	L_{ctx}	D	$S(L_{\mathrm{ctx}})$	Model	b	d	L_{ctx}	D	$S(L_{\mathrm{ctx}})$
Mistral-7B v0.1	10^4	128	32768	0	+0.056	Mistral-7B v0.3	10^6	128	32768	15	+0.242
Gemma-7B	10^4	256	8192	2	+0.069	Qwen2-7B	10^6	128	131072	9	+0.002
SmolLM2	1.3×10^5	64	8192	7	+0.079	Qwen2.5-7B	10^6	128	131072	9	+0.002
Llama-3-8B	5×10^5	128	8192	20	+0.265	Yi-6B	5×10^6	128	4096	29	+0.310
Llama-3.1-8B†	5×10^5	128	131072	16	+0.234	Granite-3.1	10^7	128	131072	17	+0.241

† Llama3 frequency scaling (Grattafiori et al., 2024) applied over base 5×10^5; all others are unscaled combs.

Operational equivalence. We compare the identity cache against the RoAM form on PG-19 perplexity and macro-averaged accuracy over seven retrieval tasks (five RULER subtasks (Hsieh et al., 2024) and two LongBench subtasks (Bai et al., 2024)) and five zero-shot tasks (LAMBADA, HellaSwag, PIQA, ARC easy and challenge). Caching in RoAM form changes perplexity by at most 0.024% (median 0.006%), against seed spreads up to 58% from resampling the documents. Per-task mean accuracies differ between the two forms by at most 0.053 (retrieval) and 0.005 (zero-shot), against single-form seed spreads up to 0.20 and 0.04; per-task results in Table 10 (Appendix F).

Per-head tuning curves. For every head we compare the measured logit versus distance curve against the prediction of Corollary 2, whose form is fixed by the spectrum and whose coefficients are the head's band cross-covariances estimated from activations (Figure 2; details in Appendix F).

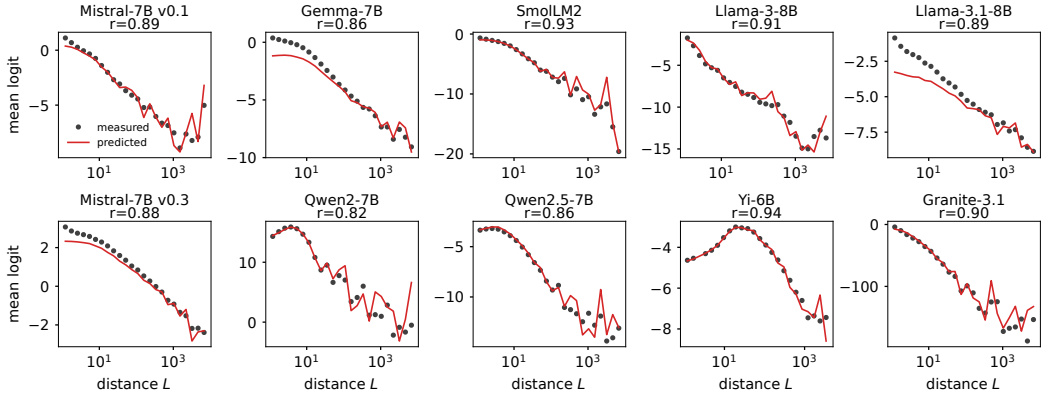


Figure 2: Per-head tuning curves: measured logit versus distance curves (points) against the predictions of Corollary 2 (lines) for each model's median head.

Centering each curve on its model-mean, the pooled median per-head correlation over 8992 live heads is 0.894 (per-model medians 0.819 to 0.938). Globally permuted predictions score $|r| \leq 0.05$, within-layer permutations 0.13 to 0.31, and every model clears its null by at least 0.62. The coefficients transfer: PG-19 cross-covariances predict WikiText-103 curves without re-estimation at a median of 0.875 (Section F.12). Fixed-lag heads (Appendix B) are under 7.4% per model, and excluding them shifts medians by at most 0.006.

Positional aliasing. For each spectrum we registered the distance at which each band set first returns within $\epsilon = 0.25$ operator deviation after exceeding it (Theorem 4; convention in Appendix F). Across ten models and eight band sets per spectrum, every onset and alias distance reproduces the registered integer, identically on all three seeds (Table 12, Appendix F). Band-sparse heads alias early: the top-two set returns within 88 to 245 tokens on every model. The top-eight and top-sixteen sets return no alias within 2×10^6 at this tolerance, certifying alias-free ranges beyond any deployed context. The shared-spectrum trio gives identical distances; Llama-3.1-8B matches every top-band integer of Llama-3-8B while differing in every low-band one, the signature of low-band-only scaling.

The long-range channel. We zero a fraction f of the spectrum, from the slow end or the fast end, on three models spanning RoPE bases 10^4 to 10^7 (Figure 3). Removing the slow half is catastrophic at every fraction. At $f = 1/8$, retrieval collapses to at most 0.005 on all three models, and perplexity rises to between 260 and 5.4×10^6. Removing the fast half is survivable. At the same fraction, perplexity stays within 9.9 and ordering holds at 0.77 to 0.85; at $f = 1/2$, with half the spectrum gone, ordering still reads 0.53 to 0.75 while the slow-side arm is destroyed by a perplexity factor of 128 to 2.9×10^4. The two arms coincide exactly at $f = 1$, where both zero the whole spectrum, in all model-seed pairs. A random-band control at matched count, separating placement from coordinates removed, is in Appendix F.6, while Appendix F.11 repeats the ablation at 65536.

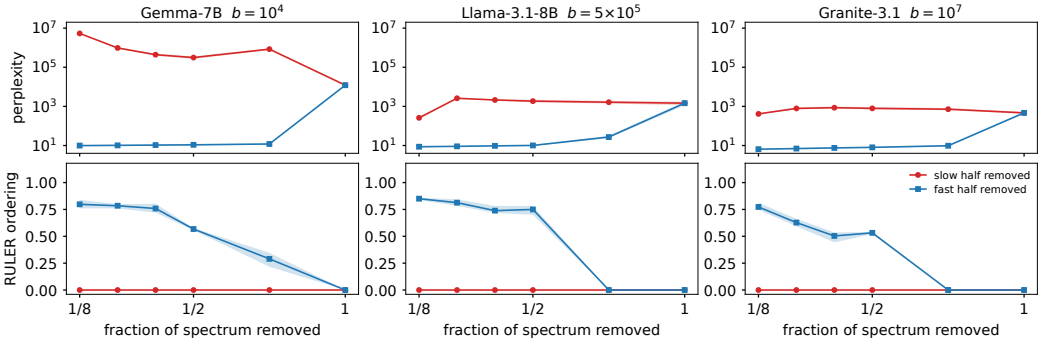


Figure 3: Band ablation: perplexity (top, log scale) and RULER ordering (bottom) against the fraction of the spectrum removed, from the slow end (red) or the fast end (blue). Lines are three-seed means, shaded bands the seed range. Removing the slow half collapses retrieval at every fraction; removing the fast half leaves ordering intact until $f = 3/4$.

Matched norm perturbation. Zeroing bands removes energy as well as position, so the previous result does not separate the two. We therefore perturb instead of remove: each stored key is edited by a δ drawn isotropically within the slow half, the fast half, or the whole spectrum, with $\|\delta_k\| = 0.25\|k\|$ per key so the certificate magnitude $\|q\|\|\delta\|$ is identical across arms. Matching is absolute, so the targeted half absorbs the whole edit: the fast arm takes $1.58\times$ to $2.08\times$ the relative perturbation of the slow arm on its own bands (Table 2), which biases the design against the prediction. Retrieval still falls further under slow-band damage on all three models, by 0.032 to 0.134 on the retrieval macro and by up to 0.356 on ordering, with the uniform arm between the two throughout. Of the 21 task-model cells, 19 order this way. Zero-shot accuracy moves by at most 0.010 while retrieval moves by 0.134, so the same perturbation that costs long-range retrieval leaves short-range prediction intact.

Table 2: Matched-norm perturbation at 0.25 of each key's norm: mean ± sample standard deviation over three seeds. Arms differ only in which half of the spectrum absorbs the edit. Relative is the mean perturbation of the targeted half; Retrieval macro-averages the five RULER and two LongBench subtasks; Zero-shot macro-averages the five zero-shot tasks.

Model	Arm	Relative	Retrieval	Ordering	Zero-shot	Perplexity
Gemma-7B	slow	0.305	0.578 ± 0.013	0.731 ± 0.007	0.646 ± 0.004	9.92 ± 0.69
	uniform	0.250	0.594 ± 0.013	0.757 ± 0.045	0.647 ± 0.004	9.88 ± 0.68
	fast	0.490	0.611 ± 0.016	0.780 ± 0.017	0.647 ± 0.004	9.83 ± 0.67
Llama-3.1-8B	slow	0.303	0.648 ± 0.024	0.634 ± 0.105	0.680 ± 0.005	8.84 ± 0.47
	uniform	0.250	0.691 ± 0.002	0.761 ± 0.041	0.682 ± 0.007	8.66 ± 0.49
	fast	0.478	0.733 ± 0.032	0.851 ± 0.066	0.682 ± 0.003	8.54 ± 0.47
Granite-3.1	slow	0.239	0.578 ± 0.012	0.501 ± 0.023	0.676 ± 0.002	6.97 ± 0.42
	uniform	0.213	0.653 ± 0.018	0.708 ± 0.069	0.679 ± 0.003	6.55 ± 0.39
	fast	0.499	0.713 ± 0.014	0.857 ± 0.030	0.686 ± 0.004	6.35 ± 0.36

Capacity frontier. We simulate the retrieval event of Theorem 3 on each deployed spectrum, with Gaussian stores of 4 to 1024 keys at offsets of up to 10^5, three seeds, and 200 trials per cell. Gaussian

stores realize the theorem's event exactly, so its explicit constants can be tested rather than fitted, and the spectrum is the only quantity that differs across models. Every cell the theory calls a success reaches a retrieval rate ≥ 0.9 and every cell it calls a failure stays at ≤ 0.5, with no violations across the 63 and 74 decided cells, and every in-grid crossing of rate 0.5 falls inside the predicted store-size bracket (Table 17, Appendix F). The remaining 63 cells lie between the two thresholds, where the theorem is silent. Appendix F.10 repeats the experiment on captured keys: the predicted margin orders the measured rate within every head on all ten models, and across heads on eight of ten.

Positional leakage in keys. The analysis treats stored keys as position-free, which trained models need not satisfy. We bound the violation on content-matched tiled prompts: the same token recurs at eight positions with identical content, so any variation in its key is position or residual context drift, and the observed variation upper-bounds η_k. The bound is 0.19 to 0.39 at the worst layer of each model, with per-model medians of 0.13 to 0.26 (Figure 9, Appendix F). The worst layer sits mid-stack in every model, between 41% and 64% of depth, never at the input or output. At this bound, the robustness shift of Theorem 6 reverses no capacity-frontier prediction.

Optimal spectra. Theorem 5 gives each spectrum's certified scale-free range L^* in closed form, the largest range whose coverage profile it dominates. Every deployed comb certifies exactly its base, and all but Mistral-7B v0.1 certify their declared contexts; llama3 scaling adds long-scale coverage while leaving L^* at the base, above the declared 131072. Matched-budget alternatives realize the theorem's exclusions. Linear-frequency schedules violate coverage by 22 to 83 bands, linear-period schedules forfeit 69% to 93% of the comb's resolution, and the best of 64 random draws reaches zero deficit at full resolution on none of the ten models (Table 3; Appendix F). The budget identity permits only these two failures: a schedule can spend coverage early and run short at distance, or spread it thin and lose the resolution endpoint, but not avoid both.

Table 3: Optimality at deployed budgets. L^*: certified range; cert.: whether it covers the declared context. Alternatives share each model's band count and endpoints; deficit: worst coverage shortfall in bands, res.: resolution relative to the comb, random: best of 64 draws; *shared spectrum.

	Mistral-7B v0.1	Gemma-7B	SmolLM2	Llama-3-8B	Llama-3.1-8B	Mistral-7B v0.3*	Qwen2-7B*	Qwen2.5-7B*	Yi-6B	Granite-3.1
L^*	10^4	10^4	1.3×10^5	5×10^5	5×10^5	10^6	10^6	10^6	5×10^6	10^7
cert.	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
lin-θ deficit	41	83	22	46	46	47	47	47	48	49
lin-period res.	0.14	0.07	0.31	0.19	0.19	0.19	0.19	0.19	0.21	0.22
random def./res.	1 / 0.79	1 / 0.76	1 / 0.77	1 / 0.99	0 / 0.76	1 / 0.74	1 / 0.76	1 / 0.84	1 / 0.85	2 / 0.95

A certificate for deployed compression. Compression acts on the memory along three axes: precision (quantization), tokens (eviction), and dimensions (low-rank projection). Proposition 3 certifies the two that edit keys in place: any perturbation δ moves any logit at any offset by at most $\|q\|\|\delta\|$, in either rotation frame. We audit four methods in their own implementations, KIVI-2, KVQuant-3, PALU, and H2O on all ten models and three seeds. Across 115 million measured (query, key, offset) triples from the three reconstruction methods, the maximum ratio is 0.93 and no per-model median tightness exceeds 0.18 (Table 4): the bound is a worst-case outer envelope, loose by a factor of 6 to 25 on the median read. H2O is classified rather than certified: retained slots are verified exact and evicted slots carry $\delta = -k$, where the bound is valid but uninformative, so the token axis is uncertifiable per instance. Rank informativeness follows error structure, 0.18 under near-uniform quantization and 0.71 under structured low-rank projection: uniform error carries little per-key rank information, while structured error carries the most.

Table 4: Certificate on deployed methods: campaign statistics (left) and quality at the per-method settings (right). Tightness and ρ are per-model ranges, ΔPPL the ten-model median, and ΔRetr and ΔZS ten-model means over the seven retrieval and five zero-shot tasks, relative to the identity cache.

Method	triples	viol.	max ratio	tightness p_{50}	ρ	ΔPPL	ΔRetr	ΔZS
KIVI-2	38,292,416	0	0.76	[0.042, 0.094]	[0.18, 0.52]	+3.7%	−0.14	0.00
KVQuant-3	38,292,416	0	0.72	[0.045, 0.097]	[0.23, 0.55]	+0.8%	−0.02	0.00
PALU	38,292,416	0	0.93	[0.039, 0.176]	[0.29, 0.71]	+61.5%†	−0.30	−0.04
H2O	classified evictor: retained slots exact; 94% of slots evicted					+10.0%	−0.51	0.00

What the certificate predicts. Across the forty (method, model) pairs, degradation is retrieval-first: the asymmetry ΔRetr − ΔZS is negative in 39 of 40 and exceeds one seed standard deviation in 36, where negative means retrieval degrades more than zero-shot accuracy (Figure 4). The one exception, KVQuant-3 on Granite-3.1, is a tie within seed noise. H2O at 94% eviction is the most extreme pair in every model and seed, taking retrieval from 0.592 to 0.078 while perplexity changes

by 10%: short-range metrics are insensitive to the removal of coherent long-range structure. Two functions survive the audit: a per-key worst-case bound computable before deployment, and a mechanism taxonomy that designates the failure mode to which standard benchmarks are insensitive.

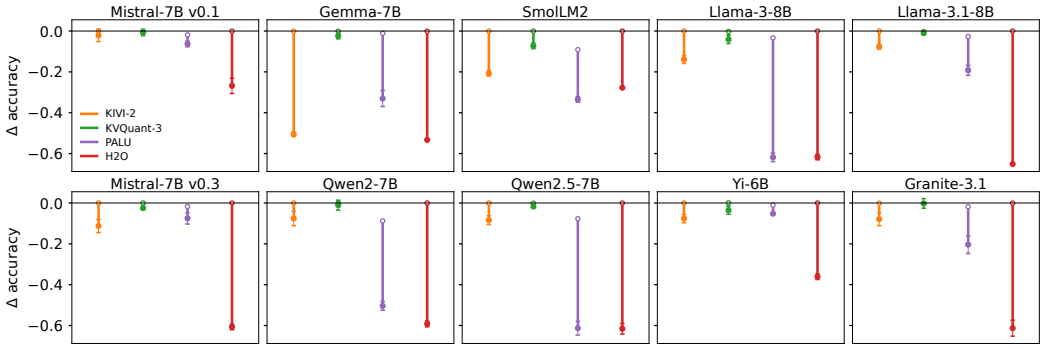


Figure 4: Accuracy degradation by model: each segment connects a method's zero-shot change (open marker) to its retrieval change (filled marker, ± 1 sample standard deviation over seeds), relative to the identity cache. The signed segment is the asymmetry, negative in 39 of 40 pairs.

8 Related Work

RoAM draws on three lines. Attention reads as associative memory under the Hopfield correspondence (Ramsauer et al., 2020; Krotov & Hopfield, 2016); rotary embeddings and their context extensions supply the positional structure (Su et al., 2024; Chen et al., 2023; Peng et al., 2024); and frequency-domain analyses characterize what makes that structure useful (Barbero et al., 2024). None supplies the per-head quantities that decide retrieval, which is what RoAM adds: a capacity law with closed-form range, aliasing distances verified to the integer, and a scale-free characterization of the deployed comb. The certificate targets empirical KV-cache compression, quantization (Liu et al., 2024; Hooper et al., 2024), eviction (Zhang et al., 2023; Xiao et al., 2024), and low-rank projection (Chang et al., 2024), all validated on benchmarks after deployment; RoAM bounds any of them per key before a task is run. FreqKV (Kai et al., 2026) also compresses in a frequency domain, but along the sequence axis; RoAM's bands are the RoPE frequencies on the head-dimension axis. Our leakage analysis reports bounds against matched permutation nulls rather than detections.

9 Conclusion

We modeled the KV cache as a Rotary Associative Memory (RoAM) to make long-context retrieval predictable and compression certifiable before deployment. We introduced theorems on retrieval range, capacity at distance, positional aliasing, and spectrum optimality, together with a position-uniform perturbation certificate for stored keys. Across ten models, predicted attention-logit curves match measured ones head by head with coefficients estimated from activations and none fitted to the curves (median detrended $r = 0.894$), the theory's constants classify the capacity frontier, all 80 pre-registered aliasing cells reproduce to the integer, and each deployed spectrum certifies a closed-form scale-free range. The certificate is audited on 115 million measured perturbations from deployed methods in their own implementations. RoAM yields a per-key bound computable before deployment and advance identification of the failure mode perplexity does not detect. The bound holds for any orthogonal positional action; an encoding that is not a unitary representation of translation, such as the additive bias of ALiBi, lies outside the frame.

Limitations. Behavioral measurements use base checkpoints, with one instruction-tuned replication (Section F.13) on which the per-head fit, the band ablation, and retrieval-first degradation under compression all carry over. Head-level measurements are estimated on PG-19; Section F.12 replicates them on WikiText-103. Retrieval is evaluated at 8192 tokens for the compression audit; Section F.11 extends retrieval and the band ablation to 65536. Mistral-7B v0.1 applies a 4096-token attention window, which bounds its retrieval independently of its spectrum. The leakage analysis bounds the position dependence of stored keys rather than detecting a positional code, though the matched-null probe resolves a small excess on WikiText-103 (Section F.12). The compression audit covers fixed-rate scalar quantization, low-rank projection, and eviction; entropy-coded and vector-quantized compressors are untested, though the certificate covers any key perturbation.

REFERENCES

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. COLM 2025.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024. ICLR 2025.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed S Abdelfattah, and Kai-Chiang Wu. Palu: Compressing kv-cache with low-rank projection. *arXiv preprint arXiv:2407.21118*, 2024. ICLR 2025.

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

IBM Granite Team. Granite 3.0 language models. *URL: https://github. com/ibm-granite/granite-3.0-language-models*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun S Shao, Kurt Keutzer, and Amir Gholami. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *Advances in Neural Information Processing Systems*, 37:1270–1303, 2024.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024. COLM 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Jushi Kai, Yixuan Wang, Boyi Zeng, Haoli Bai, Bo Jiang, Ziwei He, and Zhouhan Lin. Freqkv: Key-value compression in frequency domain for context window extension. In *International Conference on Learning Representations*, volume 2026, pp. 105568–105587, 2026.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364, 2023.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024. ICML 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. NeurIPS 2016.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1525–1534, 2016.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 31932–31951, 2024.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020. ICLR 2021.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Mingyu Xu, Xin Men, Bingning Wang, Qingyu Zhang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Base of rope bounds context length. *Advances in Neural Information Processing Systems*, 37:87386–87410, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang,

Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in neural information processing systems*, 36:34661–34710, 2023.

A Positional Structure

This appendix contains the full proofs for Section 2, the post-action metric distortion lemma, the energy view of the clock action, and the modeling dictionary.

A.1 Proof of Lemma 1

Proof. By unitarity and the homomorphism property, $\rho(t)^{-1} = \rho(-t)$ and $\langle \rho(t)\, q,\, \rho(j)\, k\rangle = \langle q,\, \rho(t)^{-1}\rho(j)\, k\rangle = \langle q,\, \rho(j-t)\, k\rangle$. Identical scores induce identical softmax weights, hence identical readouts against the same values. □

A.2 Post-action metric distortion

The pre-action frame avoids a conflation that arises in the post-action frame.

Lemma 3 (Post-action metric distortion)**.** *For all keys k, k' and positions i, j:*

$$\langle \rho(i)\, k,\, \rho(j)\, k'\rangle \;=\; \mathrm{Re}\sum_m \bar{k}_m\, k'_m\, e^{i\theta_m (j-i)}. \tag{3}$$

In particular, $\langle \rho(i)\, k,\, \rho(j)\, k\rangle = \sum_m |k_m|^2 \cos(\theta_m(j-i))$.

Proof. Expanding the inner product in the complexified basis and applying unitarity of each band rotation gives the first identity. The second follows by setting $k' = k$ and taking the real part. □

The inner product between post-action keys depends on the offset $j - i$ even when the keys are identical. Content similarity and temporal separation are therefore inseparable in this frame, which is why Lemma 1 and the pre-action frame are essential.

A.3 Proof of Proposition 1

Proof. (i) The representation is block-diagonal: $\rho(\Delta) = \bigoplus_m R(\theta_m \Delta)$ with $R(\phi)$ the 2×2 rotation by ϕ. The operator norm of the difference is the maximum over bands:

$$\|\rho(t_1) - \rho(t_2)\|_{\mathrm{op}} = \max_m \|R(\theta_m t_1) - R(\theta_m t_2)\|_{\mathrm{op}} = \max_m 2\left|\sin\Big(\frac{\theta_m (t_1 - t_2)}{2}\Big)\right|, \tag{4}$$

using $|e^{i\alpha} - e^{i\beta}| = 2|\sin((\alpha-\beta)/2)|$ on each block.

(ii) The score difference is

$$s(q, k, t_1) - s(q, k, t_2) = \mathrm{Re}\sum_m \bar{q}_m k_m \left(e^{-i\theta_m t_1} - e^{-i\theta_m t_2}\right), \tag{5}$$

and the triangle inequality applied band by band gives the stated bound with $a_m = |\bar{q}_m k_m|$. For tightness, choose the phases φ_m so that every summand $c_m(e^{-i\theta_m t_1} - e^{-i\theta_m t_2})$ is real and nonnegative; the triangle inequality is then an equality.

(iii) Let $\delta := s' - s$ with $\|\delta\|_\infty \le \varepsilon$, and set $w(u) := \mathrm{softmax}(\beta(s + u\,\delta))$ for $u \in [0, 1]$. Differentiating, $\frac{d}{du} w_i = \beta\, w_i(\delta_i - \bar{\delta})$ with $\bar{\delta} := \langle w,\, \delta\rangle$, so $\|\frac{d}{du} w\|_1 = \beta \sum_i w_i |\delta_i - \bar{\delta}|$. Split the sum by the sign of $\delta_i - \bar{\delta}$: the positive and negative parts balance by the definition of $\bar{\delta}$, each equal to some $S \ge 0$, so the sum is $2S$. With W the total weight on the positive side, $S \le W(\delta_{\max} - \bar{\delta})$ and $S \le (1 - W)(\bar{\delta} - \delta_{\min})$; multiplying and using $W(1-W) \le 1/4$ together with $(\delta_{\max} - \bar{\delta})(\bar{\delta} - \delta_{\min}) \le ((\delta_{\max} - \delta_{\min})/2)^2 \le \varepsilon^2$ gives $S \le \varepsilon/2$, hence $\|\frac{d}{du} w\|_1 \le \beta\varepsilon$ for every u. Integrating over $[0, 1]$ yields $\|w - w'\|_1 \le \beta\varepsilon$, a total-variation shift of at most $\beta\varepsilon/2$. The constant is sharp for ℓ_1 to leading order: two entries with logit shifts $\pm\epsilon$ give $\|w - w'\|_1 = \tanh(\beta\epsilon) = \beta\epsilon\,(1 + O(\beta^2\epsilon^2))$. □

A.4 Proof of Theorem 1

Proof. Multiset storage assigns entries j_1 and j_2 the same score under every query, since their values coincide. On the hypothesized band, $e^{i\theta_m (j_1 - j_2)} k_m \ne k_m$ because $k_m \ne 0$ and rotation is nontrivial, so $v := \rho(j_1)\, k - \rho(j_2)\, k \ne 0$. The query $q := \rho(-t)\, v$ separates the rotary scores: $\langle \rho(t)\, q,\, \rho(j_1)\, k\rangle - \langle \rho(t)\, q,\, \rho(j_2)\, k\rangle = \langle v,\, v\rangle = \|v\|^2 > 0$. No multiset readout reproduces this separation. □

Remark 1 (Sharpness). *If instead every band on which k is supported satisfies $\theta_m(j_1 - j_2) \in 2\pi\mathbb{Z}$, then $\rho(j_1)\,k = \rho(j_2)\,k$ and rotary attention itself assigns the two entries equal scores for every query. The hypothesis of Theorem 1 is therefore the precise obstruction.*

A.5 Proof of Theorem 2

Proof. (iv) $\Rightarrow$ (iii): on band m, both M and $\rho(\Delta)$ act as multiplication by complex scalars (λ_m and $e^{i\theta_m \Delta}$), which commute. (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i): restriction.

(i) $\Rightarrow$ (iv): complexify. The operator $\rho(1)$ is diagonal on $\mathbb{C}^d$ with eigenvalues $\{e^{i\theta_m}, e^{-i\theta_m}\}_{m=1}^{d/2}$, and these d numbers are pairwise distinct: $e^{i\theta_m} = e^{i\theta_{m'}}$ forces $\theta_m = \theta_{m'}$ because the frequencies lie in $(0, 1] \subset (0, 2\pi)$, while $e^{i\theta_m} = e^{-i\theta_{m'}}$ forces $\theta_m + \theta_{m'} \in 2\pi\mathbb{Z}$, impossible because $\theta_m + \theta_{m'} \in (0, 2]$ and $2 < 2\pi$. A matrix commuting with a diagonalizable operator of simple spectrum preserves each one-dimensional eigenspace and is therefore diagonal in the same eigenbasis. Since M is real it commutes with complex conjugation, which exchanges the two eigenlines of band m; the corresponding diagonal entries form a conjugate pair $\lambda_m, \bar{\lambda}_m$, and on the underlying real band M acts as $(\mathrm{Re}\ \lambda_m)I + (\mathrm{Im}\ \lambda_m)J$, which is multiplication by $\lambda_m \in \mathbb{C}$. □

Remark 2 (A single generator suffices; rational independence is not needed). *A natural route from integer to real shifts argues by density of the integer orbit $\{\rho(n)\}_{n\in\mathbb{Z}}$ in $U(1)^{d/2}$, which by Kronecker's theorem requires $1, \theta_1/2\pi, \ldots, \theta_{d/2}/2\pi$ to be jointly rationally independent. Deployed geometric combs violate this: for $b = 10^4$ and $d = 128$, $\theta_m = 10^{-(m-1)/16}$, so $\theta_{m+16} = \theta_m/10$ is a rational relation. The proof above needs no such hypothesis: a single integer shift has simple spectrum, and that alone forces band-diagonality.*

Remark 3 (The domain matters). *At $\theta_m = \pi$ the band rotation is $R(\pi) = -I$ and the band's eigenvalue pair collides; every real-linear map of that band then commutes with $\rho(1)$, and two such bands could mix. Distinct frequencies in $(0, \pi)$ are what the argument needs; the domain $(0, 1]$ of Section 2.1 sits inside.*

Remark 4 (Equivariant projectors are band selectors). *Idempotents in the commutant satisfy $\lambda_m^2 = \lambda_m$, hence $\lambda_m \in \{0, 1\}$: the equivariant orthogonal projectors are the band selectors.*

A.6 Energy view and clock action

In content coordinates the dense-associative-memory energy of the readout becomes time-indexed,

$$\tilde{E}_t(\xi) \;=\; -\tfrac{1}{\beta}\log\sum_s e^{\beta\langle \xi,\, \rho(s-t)\,k_s\rangle} \;+\; \tfrac{1}{2}\|\xi\|^2, \tag{6}$$

and advancing the clock acts by composition with the group. The following proposition records that the entire retrieval dynamics transports rigidly.

Proposition 4 (Clock covariance of the retrieval dynamics). *Fix a stored set $\{(k_s, s)\}$ and define the autoassociative update $T_t(\xi) = \sum_s p_s(\xi; t)\,\rho(s-t)\,k_s$ with $p_s(\xi; t) = \mathrm{softmax}_s(\beta\langle \xi,\, \rho(s-t)\,k_s\rangle)$. For every Δ: (i) $p_s(\xi;\, t+\Delta) = p_s(\rho(\Delta)\,\xi;\, t)$ for all s and ξ; (ii) $\tilde{E}_{t+\Delta} = \tilde{E}_t \circ \rho(\Delta)$ and $T_{t+\Delta} = \rho(-\Delta) \circ T_t \circ \rho(\Delta)$, hence $T_{t+\Delta}^n = \rho(-\Delta) \circ T_t^n \circ \rho(\Delta)$ for every n; (iii) ξ^* is a fixed point of T_t iff $\rho(-\Delta)\,\xi^*$ is a fixed point of $T_{t+\Delta}$; basins of attraction map isometrically under $\rho(-\Delta)$; energies correspond, $\tilde{E}_{t+\Delta}(\rho(-\Delta)\,\xi) = \tilde{E}_t(\xi)$; and the Jacobians at corresponding fixed points are similar, so their spectra and local convergence rates are invariant.*

Proof. Orthogonality of $\rho(\Delta)$ gives $\langle \xi,\, \rho(s-t-\Delta)\,k_s\rangle = \langle \rho(\Delta)\,\xi,\, \rho(s-t)\,k_s\rangle$, so the logit vectors at $(\xi, t+\Delta)$ and $(\rho(\Delta)\,\xi, t)$ coincide entrywise; this yields (i), and the same identity applied inside Equation 6 together with $\|\rho(\Delta)\,\xi\| = \|\xi\|$ yields the energy law. For the map, $\rho(s-t-\Delta)\,k_s = \rho(-\Delta)\,\rho(s-t)\,k_s$, so $T_{t+\Delta}(\xi) = \rho(-\Delta)\,T_t(\rho(\Delta)\,\xi)$; iterate for T^n. Claim (iii) follows by substitution, the energy law, and the chain rule. □

Remark 5 (Scope). *The conjugation concerns a fixed stored set; entries written between t and $t+\Delta$ extend the sum and sit outside the statement. Cache re-indexing over a fixed window rotates every attractor, basin, and rate rigidly; nothing is created, destroyed, or reshaped.*

Remark 6 (Heteroassociative readout). *Attention emits $\sum_s p_s v_s$ with values distinct from keys. Claim (i) holds for arbitrary values; the fixed-point and energy-descent reading requires the autoassociative case $v_s = \rho(s-t)\,k_s$. All score-level results of the paper concern the logits and softmax weights and are readout-agnostic.*

A.7 MODELING DICTIONARY

Table 5: Correspondence between the standard KV cache under RoPE and the rotary associative memory view. **Bold** entries mark the objects whose properties the theory characterizes.

KV cache under RoPE	Rotary associative memory
cached key k_s at position s	stored pattern with **address** $\rho(s)\ k_s$
cached value v_s	associated value of the entry
query q_t at position t	cue in the frame $\rho(t)$
attention logit $\langle \rho(t)\ q_t,\ \rho(s)\ k_s \rangle$	**retrieval score** $\langle q_t,\ \rho(-L)\ k_s \rangle, L = t - s$
attention output	softmax readout of the values
RoPE frequency configuration	the **spectrum** Θ
context-window advance	**clock action** $\tilde{E}_{t+\Delta} = \tilde{E}_t \circ \rho(\Delta)$ (Proposition 4)

B CAPACITY AT A DISTANCE

B.1 PROOF OF LEMMA 2

Proof. Split the sum $S(L) = 2\sum_m \sigma_m^2 \cos(\theta_m L)$ at $\theta_m L = 1$. Coherent terms ($\theta_m L \le 1$) satisfy $\cos(\theta_m L) \in [\cos 1, 1]$, contributing at least $2\cos(1)\, E_{\mathrm{coh}}(L)$. Incoherent terms satisfy $\cos(\theta_m L) \in [-1, 1]$, contributing at least $-2(E_{\mathrm{tot}} - E_{\mathrm{coh}}(L))$. The upper bound is $2E_{\mathrm{tot}}$ because every cosine is at most one. □

B.2 PROOF OF THEOREM 3: SETUP

Ensemble (A-Ens). $k_j = \Sigma^{1/2} g_j$, $j = 1, \dots, N$, where $g_1, \dots, g_N$ are i.i.d. with independent, mean-zero, unit-variance, K-sub-Gaussian entries ($\|g_{j,i}\|_{\psi_2} \le K$); $\Sigma \succeq 0$ with $\mathrm{Tr}\,\Sigma = 1$. Hanson–Wright requires independent coordinates, which "sub-Gaussian with covariance Σ" alone does not supply; the linear-generative form ensures this. Gaussian keys correspond to $g \sim \mathcal{N}(0, I)$, $K = 1$.

Scores. Offsets $L_1, \dots, L_N \ge 0$ are fixed before the keys are drawn (arbitrary, possibly adversarial, not key-dependent). The target k_J sits at distance $L_J = L$, the query is the content echo $q = k_J$, and $s_j = \langle q,\ \rho(-L_j)\, k_j \rangle$ per Equation 1. The sign of the offset is immaterial below: the signal depends on $\rho(-L)$ only through its symmetric part D_L, and the distractor analysis uses only the orthogonality of ρ. Retrieval succeeds at level α when the softmax weight $a_J \ge 1 - \alpha$.

Spectral quantities.

$$S(L) := \mathrm{Tr}(D_L \Sigma) = 2\sum_m \sigma_m^2 \cos(\theta_m L), \tag{7}$$

$$V_{\mathrm{sig}}(L) := \|\Sigma^{1/2} D_L \Sigma^{1/2}\|_F^2, \quad v_*(L) := \|\Sigma^{1/2} D_L \Sigma^{1/2}\|_{\mathrm{op}} \le \|\Sigma\|_{\mathrm{op}}, \tag{8}$$

$$V_{\mathrm{int}} := \sup_{\Delta} \mathrm{Tr}\big(\rho(\Delta)^\top \Sigma\, \rho(\Delta)\, \Sigma\big). \tag{9}$$

D_L is band-diagonal, so the identity for $S(L)$ holds for arbitrary cross-band correlation. In general $V_{\mathrm{sig}} \le V_{\mathrm{int}} \le \|\Sigma\|_{\mathrm{op}}\, \mathrm{Tr}\,\Sigma = \|\Sigma\|_{\mathrm{op}}$: for the first inequality, write $B := \Sigma^{1/2}\rho(L)\Sigma^{1/2}$, so that $\Sigma^{1/2} D_L \Sigma^{1/2} = \frac{1}{2}(B + B^\top)$ and $V_{\mathrm{sig}} \le \|B\|_F^2 = \mathrm{Tr}(\rho(L)^\top \Sigma \rho(L) \Sigma) \le V_{\mathrm{int}}$; the second is $\mathrm{Tr}(AB) \le \|A\|_{\mathrm{op}}\, \mathrm{Tr}\, B$ for positive semidefinite A, B. For band-isotropic $\Sigma = \bigoplus_m \sigma_m^2 I_2$, $\rho(\Delta)\Sigma\rho(\Delta)^\top = \Sigma$ for all Δ, giving $V_{\mathrm{int}} = \mathrm{Tr}(\Sigma^2) = 2\sum_m \sigma_m^4$, independent of the offsets. For reference in the results: with $\bar{V} := \sum_m \sigma_m^4$, band-isotropically $V_{\mathrm{int}} = 2\bar{V}$ and the simulation threshold $\sqrt{4\bar{V}\log N}$ equals $\sqrt{2V_{\mathrm{int}}\log N}$.

B.3 SIGNAL CONCENTRATION

Lemma 4. *Under (A-Ens),* $\Pr\big(|s_J - S(L)| \ge u\big) \le 2\exp\big(-c_{\mathrm{hw}} \min(u^2/K^4 V_{\mathrm{sig}},\ u/K^2 v_*)\big)$. *With*

$$u_A(\delta) := K^2 \sqrt{V_{\mathrm{sig}} \log(2/\delta)/c_{\mathrm{hw}}} + K^2 v_* \log(2/\delta)/c_{\mathrm{hw}}, \tag{10}$$

$\Pr(|s_J - S(L)| \ge u_A(\delta)) \le \delta$.

Proof. For any antisymmetric A, $x^\top Ax = 0$, so $s_J = k_J^\top D_L k_J = g_J^\top M g_J$ with $M := \Sigma^{1/2} D_L \Sigma^{1/2}$ symmetric. The expectation is $\mathbb{E}[g^\top Mg] = \operatorname{Tr} M = \operatorname{Tr}(D_L\Sigma) = S(L)$. The entries of g_J are independent, mean-zero, and K-sub-Gaussian, so Hanson–Wright (Table 6, row HW) applies with $\|M\|_F^2 = V_{\mathrm{sig}}$ and $\|M\|_{\mathrm{op}} = v_*$. For the inversion, if u exceeds both summands of $u_A(\delta)$ then $c_{\mathrm{hw}} \min(\cdot) \geq \log(2/\delta)$; their sum dominates both, so $u = u_A(\delta)$ suffices. □

Explicit constants for Gaussian keys. For $g \sim \mathcal{N}(0, I)$ and symmetric M with spectrum λ, split $\lambda = \lambda_+ - \lambda_-$ and apply the Laurent–Massart bounds (Table 6, row LM) to each part. Since $\{Z \geq u_+ + u_-\} \subseteq \{Z_+ \geq u_+\} \cup \{-Z_- \geq u_-\}$ with $Z_\pm := \sum_i \lambda_{\pm,i}(g_i^2 - 1)$, and $\|\lambda_+\|_2 + \|\lambda_-\|_2 \leq \sqrt{2}\,\|M\|_F$, $\|\lambda_+\|_\infty \leq \|M\|_{\mathrm{op}}$:

$$\Pr\Big(|g^\top Mg - \operatorname{Tr} M| \geq 2\sqrt{2}\,\|M\|_F \sqrt{t} + 2\|M\|_{\mathrm{op}}\, t\Big) \leq 4e^{-t} \quad \text{for all } t > 0. \tag{11}$$

Lemma 4 therefore holds for Gaussian keys with the fully explicit

$$u_A^G(\delta) := 2\sqrt{2}\,\sqrt{V_{\mathrm{sig}} \log(4/\delta)} + 2\, v_* \log(4/\delta), \tag{12}$$

with no unspecified constant. Every Gaussian-ensemble consequence in this appendix, including the failure clause, sandwich below, and simulation thresholds of the results, can be run with u_A^G in place of u_A, removing c_{hw} from the empirical interface; c_{hw} remains only in the sub-Gaussian generality.

B.4 Norm concentration

Lemma 5. *Under (A-Ens), $\|k_J\|^2 = g_J^\top \Sigma g_J$ has mean 1 and satisfies the same tail as Lemma 4 with $(\|\Sigma\|_F^2, \|\Sigma\|_{\mathrm{op}})$ in place of (V_{sig}, v_*).*

Proof. Apply Lemma 4 with $M = \Sigma$. □

Define $u_B(\delta)$ by the corresponding inversion. Condition (N1) requires $u_B(\delta/3) \leq 1$, ensuring $\|k_J\|^2 \leq 2$ on the event. For deployed profiles, $\|\Sigma\|_F^2 = O(1/d)$, so (N1) holds by orders of magnitude.

B.5 Interference

Lemma 6. *Conditioned on k_J, the distractor scores $\{s_j\}_{j \neq J}$ are conditionally independent, mean zero, with*

$$\operatorname{Var}(s_j \mid k_J) = k_J^\top \rho(L_j)^\top\, \Sigma\, \rho(L_j)\, k_J \leq \|\Sigma\|_{\mathrm{op}}\, \|k_J\|^2, \qquad \|s_j\|_{\psi_2 | k_J} \leq C_2 K\, \|\Sigma\|_{\mathrm{op}}^{1/2} \|k_J\|. \tag{13}$$

Proof. Write $s_j = \langle q, \rho(-L_j)\,\Sigma^{1/2} g_j\rangle = \langle a_j, g_j\rangle$ with $a_j := \Sigma^{1/2}\rho(L_j)\,k_J$ fixed given k_J. Independence of g_j from k_J and from each other gives the conditional structure. The variance is $\|a_j\|^2 = k_J^\top \rho(L_j)^\top \Sigma\, \rho(L_j)\, k_J$. General Hoeffding (Table 6, row GH) gives $\|\langle a, g\rangle\|_{\psi_2} \leq C_2 K \|a\|$, and orthogonality of ρ gives $\|a_j\| \leq \|\Sigma\|_{\mathrm{op}}^{1/2} \|k_J\|$. □

The uniform bound is offset-independent: adversarially placed but key-independent offsets do not affect the above; offsets chosen after observing the keys sit outside (A-Ens) and outside the theorem. For band-isotropic Σ, the conditional variance reduces to $k_J^\top \Sigma k_J$ for every offset.

B.6 Interference maximum

Lemma 7. *For $X_1, \ldots, X_M$ independent, conditionally mean-zero with $\|X_j\|_{\psi_2} \leq \tau$: $\Pr(\max_j X_j \geq \tau\sqrt{\log(2M/\delta)/c_5}) \leq \delta$.*

Proof. Sub-Gaussian tail (Table 6, row SG) and a union bound. □

B.7 SOFTMAX MARGIN

Lemma 8. *Let $\gamma := s_J - \max_{j \neq J} s_j$. Then $a_J \geq 1-\alpha$ if $\gamma \geq \beta^{-1} \log \frac{(N-1)(1-\alpha)}{\alpha}$, and $a_J \geq 1-\alpha$ iff $\sum_{j \neq J} e^{\beta(s_j - s_J)} \leq \frac{\alpha}{1-\alpha}$.*

Proof. $a_J = (1 + \sum_{j \neq J} e^{\beta(s_j - s_J)})^{-1}$ gives the equivalence. The sum is at most $(N-1)e^{-\beta\gamma}$, giving the sufficient condition. □

B.8 PROOF OF THEOREM 3: SUCCESS

Proof. Split δ into thirds.

On E_A (Lemma 4 at level $\delta/3$): $s_J \geq S(L) - u_A(\delta/3)$.

On E_B (Lemma 5 at $\delta/3$, using (N1)): $\|k_J\|^2 \leq 2$, so the conditional ψ_2 norm from Lemma 6 is at most $\tau := \sqrt{2}\, C_2 K \|\Sigma\|_{\mathrm{op}}^{1/2}$.

On E_B, Lemma 7 at level $\delta/3$ with $M = N-1$ gives $\max_{j \neq J} s_j \leq \tau\sqrt{\log(6N/\delta)/c_5}$ with conditional probability at least $1 - \delta/3$; call this event E_D. Integrating over $k_J \in E_B$ keeps the unconditional failure at most $\delta/3$.

On $E_A \cap E_B \cap E_D$, the margin satisfies $\gamma \geq S(L) - u_A(\delta/3) - \tau\sqrt{\log(6N/\delta)/c_5}$, which the hypothesis makes at least the Lemma 8 threshold. □

Corollary 8 (From the explicit form to the main-text constant)**.** *For $\alpha = \delta \leq 1/3$, the success condition of Theorem 3 holds with $C = \max\big(3/\sqrt{c_{\mathrm{hw}}},\ 3\sqrt{2}\, C_2/\sqrt{c_5},\ 3/c_{\mathrm{hw}},\ 3\big)$, absorbing $u_A(\delta/3)$, the interference term, and the margin term into the four displayed summands. The second summand dominates the first when $\|\Sigma\|_{\mathrm{op}} \geq V_{\mathrm{sig}} K^2 \log(1/\delta)/\log(N/\delta)$, which holds in the flat case at moderate δ; the $v_* \log(1/\delta)$ summand is the sub-exponential Hanson–Wright branch and is lower-order outside extreme-δ regimes.*

Corollary 9 (Band-isotropic sandwich; fully explicit for Gaussian keys)**.** *Assume Gaussian keys, band-isotropic Σ, $r_4 := (\operatorname{Tr}\Sigma^2)^2/\operatorname{Tr}\Sigma^4 \geq 16\log(4/\delta)$, and $\|\Sigma\|_{\mathrm{op}}^2 \log(4/\delta) \leq V_{\mathrm{int}}/4$. Then success at level α holds with probability at least $1-\delta$ whenever $S(L) \geq u_A^G(\delta/4) + 2\sqrt{V_{\mathrm{int}} \log(4N/\delta)} + \beta^{-1} \log \frac{(N-1)(1-\alpha)}{\alpha}$, and failure holds as in Theorem 3 with u_A^G. The crossover kernel value at which retrieval flips is therefore pinned between $\frac{1}{2}\sqrt{V_{\mathrm{int}} \log(N-1)}$ and $2\sqrt{V_{\mathrm{int}} \log(4N/\delta)}$ up to the lower-order u_A^G and β^{-1} terms: a factor-$4(1+o(1))$ sandwich with no unspecified constant.*

Proof. The Laurent–Massart upper tail on $v := k_J^\top \Sigma k_J = g^\top \Sigma^2 g$ at $t = \log(4/\delta)$ gives $v \leq V_{\mathrm{int}}\big(1 + 2\sqrt{t/r_4}\big) + 2\|\Sigma\|_{\mathrm{op}}^2 t \leq \frac{3}{2}V_{\mathrm{int}} + \frac{1}{2}V_{\mathrm{int}} = 2V_{\mathrm{int}}$ under the stated conditions, except with probability $\delta/4$. Conditionally on k_J, the distractors are i.i.d. $\mathcal{N}(0, v)$ (Lemma 6 and band-isotropy), so a Gaussian tail with a union bound gives $\max_{j \neq J} s_j \leq \sqrt{2v \log(4N/\delta)} \leq 2\sqrt{V_{\mathrm{int}} \log(4N/\delta)}$ except with probability $\delta/4$. The signal lower bound is Equation 11 at level $\delta/4$; Lemma 8 converts the margin, and the remaining $\delta/4$ is slack. □

B.9 PROOF OF THEOREM 3: FAILURE

Two additional ingredients are needed for the converse.

Lemma 9 (Interference-variance lower bound)**.** *For Gaussian keys and band-isotropic Σ with all $\sigma_m^2 > 0$, $v := k_J^\top \Sigma k_J = g^\top \Sigma^2 g$ has mean $V_{\mathrm{int}} = \operatorname{Tr}\Sigma^2$, and the Laurent–Massart lower tail (Table 6, row LM) gives $\Pr(v \leq \frac{1}{2}V_{\mathrm{int}}) \leq \exp(-r_4/16)$ with $r_4 = (\operatorname{Tr}\Sigma^2)^2/\operatorname{Tr}\Sigma^4$. The spread condition $r_4 \geq 16\log(2/\delta)$ of Theorem 3 makes this at most $\delta/2$.*

Lemma 10 (Gaussian maximum lower bound; explicit)**.** *For $M \geq 8$ and $Z_1, \ldots, Z_M$ i.i.d. $\mathcal{N}(0, v)$:*

$$\Pr\Big(\max_j Z_j \leq \tfrac{1}{2}\sqrt{2v \log M}\Big) \ \leq\ \exp\Big(-\frac{M^{3/4}}{2\sqrt{\pi \log M}}\Big). \tag{14}$$

Proof. $\Pr(\max \leq t) = (1 - Q(t/\sqrt{v}))^M \leq \exp(-M\, Q(t/\sqrt{v}))$ with Q the standard normal upper tail. At $t = \frac{1}{2}\sqrt{2v \log M}$, $s := t/\sqrt{v} = \frac{1}{2}\sqrt{2\log M} \geq 1$ because $M \geq 8 > e^2$. The Mills bound

(Table 6, row GT) gives $Q(s) \geq \frac{s}{1+s^2}\varphi(s) \geq \frac{\varphi(s)}{2s}$, and $\varphi(s) = e^{-s^2/2}/\sqrt{2\pi} = M^{-1/4}/\sqrt{2\pi}$. Hence $M\,Q(s) \geq M^{3/4}/(2s\sqrt{2\pi}) = M^{3/4}/(2\sqrt{\pi \log M})$. □

Proof of Theorem 3, failure. Condition on k_J in the event $\{v \geq \frac{1}{2}V_{\text{int}}\}$ of Lemma 9 (probability at least $1 - \delta/2$). Given k_J, the $N-1$ distractor scores are i.i.d. $\mathcal{N}(0, v)$: the g_j are Gaussian, and band-isotropy makes the conditional variance equal to $k_J^\top \Sigma k_J = v$ for every offset. Lemma 10 with $M = N - 1 \geq 8$ gives $\max_{j \neq J} s_j > \frac{1}{2}\sqrt{2v \log(N-1)} \geq \frac{1}{2}\sqrt{V_{\text{int}} \log(N-1)}$ except with the stated stretched-exponential probability. Lemma 4 at level $\delta/2$ bounds $s_J \leq S(L) + u_A(\delta/2)$. Under the hypothesis the maximum exceeds the signal, so the argmax is a distractor and $a_J < 1/2$ by the equivalence of Lemma 8 at $\alpha = 1/2$. The total failure probability is at most $\delta/2 + \delta/2 + \exp\big(-(N-1)^{3/4}/(2\sqrt{\pi \log(N-1)})\big)$. □

Symmetric non-Gaussian ensembles. If the entries of g are symmetric with kurtosis at most κ, set $\bar{\kappa} := \max(\kappa, 3)$; then $\mathbb{E}s_j^4 = \sum_i a_i^4(\mathbb{E}g_i^4 - 3) + 3(\sum_i a_i^2)^2 \leq \bar{\kappa}v^2$, and Paley–Zygmund at level $1/4$ with the symmetry of s_j gives $\Pr(s_j \geq \frac{1}{2}\sqrt{v} \mid k_J) \geq \frac{9}{32\bar{\kappa}}$. Hence $\max_{j \neq J} s_j \geq \frac{1}{2}\sqrt{v}$ with probability at least $1 - (1 - \frac{9}{32\bar{\kappa}})^{N-1}$: a failure threshold at $\Theta(\sqrt{V_{\text{int}}})$, without the $\sqrt{\log N}$ factor, for symmetric finite-kurtosis keys. We do not claim the full $\sqrt{\log N}$ converse outside the Gaussian case.

General profiles: offset-uniform variance control. The success clause scales with $\|\Sigma\|_{\text{op}}^{1/2}$ while the failure clause scales with $V_{\text{int}}^{1/2}$, and $\operatorname{Tr}\Sigma^2 \leq \|\Sigma\|_{\text{op}}$ always (with equality iff the nonzero spectrum of Σ is flat), so for spiked profiles the two proven thresholds separate. For Gaussian keys the gap narrows. Write $A_j := \rho(L_j)^\top \Sigma \rho(L_j)$ and $v_j := k_J^\top A_j k_J = g^\top M_j g$ with $M_j := \Sigma^{1/2} A_j \Sigma^{1/2}$, so that $\mathbb{E}v_j = \operatorname{Tr}(A_j\Sigma) \leq V_{\text{int}}$, $\|M_j\|_F \leq \|\Sigma\|_{\text{op}}\|\Sigma\|_F$ (orthogonal invariance of the Frobenius norm), and $\|M_j\|_{\text{op}} \leq \|\Sigma\|_{\text{op}}^2$. Applying Equation 11 to each of the at most N matrices M_j with a union bound: with probability at least $1 - \delta/4$,

$$\max_j v_j \;\leq\; V_{\max}(\delta) \;:=\; V_{\text{int}} + 2\sqrt{2}\,\|\Sigma\|_{\text{op}}\|\Sigma\|_F\sqrt{\log(8N/\delta)} + 2\|\Sigma\|_{\text{op}}^2 \log(8N/\delta). \tag{15}$$

Conditionally on k_J in this event the distractors are independent Gaussians with variances $v_j \leq V_{\max}(\delta)$, so $\max_{j \neq J} s_j \leq \sqrt{2V_{\max}(\delta)\log(4N/\delta)}$ except with probability $\delta/4$, and success holds whenever $S(L) \geq u_A^G(\delta/4) + \sqrt{2V_{\max}(\delta)\log(4N/\delta)} + \beta^{-1}\log\frac{(N-1)(1-\alpha)}{\alpha}$. Since $V_{\text{int}} \geq \operatorname{Tr}\Sigma^2$ always (the supremum includes $\Delta = 0$), the leading interference scale improves from $\|\Sigma\|_{\text{op}}^{1/2}$ to $V_{\text{int}}^{1/2}$, meeting the failure scale up to the constant $2\sqrt{2}$ and the stated corrections. What remains open: the correction terms (logarithmic inflation for moderately spiked profiles), the failure side for non-band-isotropic Σ, and the sub-Gaussian analogue. In the band-isotropic case $v_j \equiv k_J^\top \Sigma k_J$ and no union is needed; this generalizes the interference step of Corollary 9.

B.10 Proofs of the flat-comb consequences

Proof of Corollary 2. In band coordinates $z_m = v_{2m-1} + iv_{2m}$, Equation 1 reads $s(q,k,L) = \operatorname{Re}\sum_m \bar{q}_m k_m e^{-i\theta_m L}$. The rotation is deterministic, so taking the expectation over the activation distribution at fixed L and writing $\hat{C}_m := \mathbb{E}[\bar{q}_m k_m]$ gives $\mathbb{E}\,s(q,k,L) = \operatorname{Re}\sum_m \hat{C}_m e^{-i\theta_m L} = \sum_m \big[\operatorname{Re}\hat{C}_m \cos(\theta_m L) + \operatorname{Im}\hat{C}_m \sin(\theta_m L)\big]$. No independence between q and k is required and no distributional assumption beyond the existence of the per-band second moments; $\hat{C}_m = 2\sigma_m^2$ real recovers Equation 7. □

Proof of Equation 2 and Corollary 1. For the flat comb, $\sigma_m^2 \equiv 1/d$ and $\theta_{(k)} = b^{-2k/d}$, so band-isotropy gives $V_{\text{int}} = 2\sum_m \sigma_m^4 = 1/d$, and the coherent count is $D(\Theta; L) = \#\{m : \theta_m L \leq 1\} = \frac{d}{2}(1 - \log_b L) + O(1)$. By Lemma 2 the coherent bands contribute $\frac{2\cos 1}{d}\,D(\Theta;L)$ and the incoherent terms cancel to $O(1/d)$ away from the aliases of Theorem 4, so $S(L) \asymp \frac{2}{d}D(\Theta;L)$ on the envelope. Both clauses of Theorem 3 place the crossover kernel value at $\Theta(\sqrt{V_{\text{int}}\log N}) = \Theta(\sqrt{\log N/d})$, pinned within a factor of four by Corollary 9. Equating, retrieval at (L, N) requires $\frac{2}{d}D(\Theta;L) \geq \Theta(\sqrt{\log N/d})$, that is $\log N_{\max}(L) \asymp D(\Theta;L)^2/d$ up to an absolute constant, which is Equation 2. Substituting $D(\Theta;L) = \frac{d}{2}(1 - \log_b L) + O(1)$ and solving for the offset gives $\log_b L \leq 1 - \Theta(\sqrt{\log N/d})$, which is Corollary 1. The softmax term $\beta^{-1}\log\frac{(N-1)(1-\alpha)}{\alpha}$ is nonbinding throughout, as stated with Equation 2. □

B.11 Remarks

Duplicates. A duplicate $k_{j'} = k_J$ stored at distance L' has score concentrating at $S(L')$. The target-duplicate gap concentrates at $S(L) - S(L') = -4\sum_m \sigma_m^2 \sin(\theta_m(L + L')/2)\,\sin(\theta_m(L - L')/2)$, is controlled by Proposition 1, and vanishes in particular when $L - L'$ is an alias in the sense of Theorem 4 restricted to the support of the profile. The converse fails: the reflection family $\theta_m(L + L') \in 2\pi\mathbb{Z}$ on the support also annihilates the gap with no alias of the difference (single band $\theta = 2\pi/10$, $L = 3$, $L' = 7$), and accidental cross-band cancellation can too.

Phase pre-coding. The failure clause applies to the content-echo ensemble. A head may set $\varphi_m = \theta_m L_0$, so that every band term $\cos(\theta_m L - \varphi_m)$ peaks at $L = L_0$, and thereby use all $d/2$ bands at a single lag; fixed-offset heads (previous-token, induction-style) are therefore inexpensive, while content retrieval at arbitrary distance pays the $S(L)$ price.

Nonmonotonicity. $S(L)$ is almost-periodic rather than monotone: it exhibits exceptional returns toward $2E_{\mathrm{tot}}$ at distances that are aliases of Theorem 4 restricted to the support of the profile. The working law Equation 2 describes the typical envelope; Theorem 4 characterizes the exceptional set.

Correlated distractors. The $\log N$ terms arise as the metric entropy of an i.i.d. score set. For clustered or correlated distractors, Dudley's entropy integral replaces $\log N$ by the entropy of $\{\rho(-L_j)k_j\}$, an effective count $N_{\mathrm{eff}} \le N$.

B.12 Constants

Table 6: Named constants and sources. Rows above the rule are absolute constants from the cited sources; rows below are fully explicit in this appendix. c_{hw} is required only for the sub-Gaussian generality: all Gaussian-ensemble thresholds run on the explicit u_A^G of Equation 12.

Symbol	Statement	Source
c_{hw}	Hanson–Wright inequality	Vershynin, *HDP*, Thm 6.2.1
C_2	General Hoeffding	Vershynin, Thm 2.6.3
c_5	Sub-Gaussian tail	Vershynin, Prop 2.5.2
LM	Laurent–Massart tails, $\Pr(\sum_i \lambda_i(g_i^2 - 1) \ge 2\|\lambda\|_2\sqrt{t} + 2\|\lambda\|_\infty t) \le e^{-t}$, $\Pr(\sum_i \lambda_i(g_i^2 - 1) \le -2\|\lambda\|_2\sqrt{t}) \le e^{-t}$, $\lambda_i \ge 0$	Laurent–Massart (2000), Lemma 1
GT	Mills bound $Q(s) \ge \frac{s}{1+s^2}\varphi(s)$	Standard
C	Corollary 8: $\max(3/\sqrt{c_{\mathrm{hw}}},\ 3\sqrt{2}C_2/\sqrt{c_5},\ 3/c_{\mathrm{hw}},\ 3)$, valid for $\delta \le 1/3$	this appendix
$1/(2\sqrt{\pi})$	Lemma 10 constant, hypothesis $M \ge 8$	this appendix
$1/2$	failure-threshold coefficient of Theorem 3	this appendix
u_A^G	Gaussian signal deviation Equation 12	this appendix

C Aliasing and Resolution

C.1 Proof of Theorem 4

Proof. Fix an integer $M \ge 1$ and set $n = d/2$ for part (i), $n = r$ on the active bands for part (ii). Partition $[0,1)^n$ into M^n congruent boxes of side $1/M$. Among the $M^n + 1$ points $x^{(\Delta)} := (\Delta\theta_m/2\pi \bmod 1)_m$ for $\Delta = 0, 1, \ldots, M^n$, two fall in the same box, say at $\Delta_1 < \Delta_2$. Set $\Delta^* := \Delta_2 - \Delta_1 \in [1, M^n]$. For each coordinate, the real difference $x_m^{(\Delta_2)} - x_m^{(\Delta_1)}$ lies in $(-1/M, 1/M)$ and is congruent to $\Delta^*\theta_m/2\pi$ modulo one, so $\|\Delta^*\theta_m/2\pi\| \le 1/M$ for every m, where $\|\cdot\|$ is distance to the nearest integer.

The representation is block-diagonal, so $\|\rho(\Delta^*) - I\|_{\mathrm{op}} = \max_m |e^{i\theta_m\Delta^*} - 1| = \max_m 2\,|\sin(\theta_m\Delta^*/2)|$. Writing $x_m = \Delta^*\theta_m/2\pi$, the periodicity and symmetry of the sine give $|\sin(\pi x_m)| = \sin(\pi\|x_m\|)$ (note $\|x_m\| \le 1/2$), hence

$$\|\rho(\Delta^*) - I\|_{\mathrm{op}} \;=\; \max_m 2\sin\bigl(\pi\|x_m\|\bigr) \;\le\; 2\pi \max_m \|x_m\| \;\le\; \frac{2\pi}{M}. \tag{16}$$

Taking $M = \lceil 2\pi/\varepsilon \rceil$ gives $\|\rho(\Delta^*) - I\|_{\mathrm{op}} \le \varepsilon$ with $\Delta^* \le \lceil 2\pi/\varepsilon \rceil^n$, which is part (i) for $n = d/2$ and part (ii) for the action restricted to the r active bands.

For the score consequence, $s(q,k,L+\Delta^*) - s(q,k,L) = \langle q,\, \rho(-L)\,(\rho(-\Delta^*) - I)\,k\rangle$ by Equation 1, so by orthogonality of $\rho(-L)$ and Cauchy–Schwarz, $|s(q,k,L+\Delta^*) - s(q,k,L)| \le \|\rho(-\Delta^*) - I\|_{\mathrm{op}}\,\|q\|\,\|k\| = \|\rho(\Delta^*) - I\|_{\mathrm{op}}\,\|q\|\,\|k\| \le \varepsilon\|q\|\|k\|$, using $\rho(-\Delta) - I = (\rho(\Delta) - I)^\top$. For part (ii), a pair whose coefficients c_m vanish off the support runs the same chain in the restricted subspace, with the restricted operator norm and the restricted vector norms. □

The bound is an existence guarantee, not an estimate: the first alias of a concrete spectrum can occur far below $\lceil 2\pi/\varepsilon\rceil^r$, which is why the certification search of the next subsection, not the bound, produces the reported ranges.

C.2 Certification by lattice search

Whether a given Δ is an ε-alias of the (restricted) action is a closed-form test. By the identity in the proof above,

$$\|\rho(\Delta) - I\|_{\mathrm{op}} \le \varepsilon \iff \max_{m\in\mathrm{supp}} \|\Delta\theta_m/2\pi\| \le \tfrac{1}{\pi}\arcsin(\varepsilon/2), \tag{17}$$

valid for $\varepsilon \in (0,2]$ since $\sin(\pi y)$ is increasing on $[0,1/2]$. The two-sided bounds $2y \le \sin(\pi y) \le \pi y$ on $[0,1/2]$ sandwich the exact threshold: $\varepsilon/(2\pi) \le \frac{1}{\pi}\arcsin(\varepsilon/2) \le \varepsilon/4$. Consequently a hit in $\{\Delta : \max_{m\in\mathrm{supp}} \|\Delta\theta_m/2\pi\| \le \varepsilon/(2\pi)\}$ already exhibits an ε-alias, while emptiness of $\{\Delta \le T : \max_{m\in\mathrm{supp}} \|\Delta\theta_m/2\pi\| \le \varepsilon/4\}$ certifies that no ε-alias exists below range T; the exact criterion Equation 17 decides the strip between the two thresholds. Certification of an alias-free range therefore requires the outer threshold $\varepsilon/4$ or the exact criterion; the inner threshold $\varepsilon/(2\pi)$ is sufficient only for finding aliases. A scan of $\Delta = 1,\dots,T$ against Equation 17 is the finite computation referenced in Section 4, and per-head certificates use the head's measured support.

C.3 Proof of Proposition 2

Proof. (i) For a Gaussian observation $y = s(L) + \epsilon$ with $\epsilon \sim \mathcal{N}(0,\tau^2)$, the Fisher information for the location parameter L is $I(L) = s'(L)^2/\tau^2$. Differentiating the tuning curve, $s'(L) = -\sum_m a_m\theta_m \sin(\theta_m L - \varphi_m)$. With independent uniform phases, $\mathbb{E}_\varphi \sin(\theta_m L - \varphi_m) = 0$ and $\mathbb{E}_\varphi \sin^2(\theta_m L - \varphi_m) = 1/2$, so the cross terms vanish by independence and $\mathbb{E}_\varphi\, s'(L)^2 = \frac{1}{2}\sum_m a_m^2\theta_m^2$, independent of L.

(ii) Split $\sum_m a_m^2\theta_m^2$ at $\theta_m \le 1/L$. Coherent terms satisfy $a_m^2\theta_m^2 \le a_m^2/L^2$; incoherent terms satisfy $a_m^2\theta_m^2 \le a_m^2$ because $\theta_m \le 1$. Hence $2\tau^2\,\mathbb{E}_\varphi I \le E^a_{\mathrm{coh}}(L)/L^2 + \big(E_a - E^a_{\mathrm{coh}}(L)\big) = E_a - (1 - L^{-2})\,E^a_{\mathrm{coh}}(L)$, which rearranges to the stated inequality. Equality requires $\theta_m = 1/L$ for every coherent band with $a_m > 0$ and $\theta_m = 1$ for every incoherent band with $a_m > 0$, so the pair's spectrum is supported on $\{1/L,\, 1\}$; conversely, any pair supported there attains both bounds, since $\theta = 1/L$ lies in the coherent set ($\theta L = 1$). □

Resolution functionals. For any fixed amplitude profile, the phase-averaged Fisher information of Proposition 2(i) is increasing in each frequency wherever $a_m > 0$; these are the resolution functionals referenced by the optimality result of Section 5.

D Optimal Spectra

D.1 Setting and notation

Spectra are $\Theta = \{\theta_1,\dots,\theta_{d/2}\} \subset (0,1]$ with coverage $D(\Theta;L) = \#\{m : \theta_m \le 1/L\}$, range $L_{\max} > 1$, and profile $g(L) = \frac{d}{2}\big(1 - \frac{\ln L}{\ln L_{\max}}\big)$ as in Section 5. Write the order statistics in ascending order, $\theta_{(1)} \le \cdots \le \theta_{(d/2)}$. Define the ladder $L_k := L_{\max}^{1-2k/d}$ and ceilings $c_k := 1/L_k = L_{\max}^{-(1-2k/d)}$ for $k = 1,\dots,d/2$; note $g(L_k) = k$ and $c_{d/2} = 1$. Let $\mathcal{F}$ denote the set of spectra dominating g, that is, $D(\Theta;L) \ge \lfloor g(L)\rfloor$ for all $L \in [1, L_{\max}]$.

D.2 Proof of Theorem 5(i): the budget identity

Proof. For each m, $\int_1^{L_{\max}} \mathbf{1}[\theta_m \le 1/L]\, d\ln L$ is the length of $\{\ell \in [0, \ln L_{\max}] : \ell \le \ln(1/\theta_m)\}$, which equals $\min(\ln(1/\theta_m),\, \ln L_{\max})$ because $\theta_m \le 1$ makes $\ln(1/\theta_m) \ge 0$. Summing over m gives

the identity. Each term is at most $\ln L_{\max}$, with equality iff $\theta_m \le 1/L_{\max}$; hence the bound $\frac{d}{2}\ln L_{\max}$, with equality iff every frequency is at most $1/L_{\max}$. □

D.3 Proof of Theorem 5(ii): the characterization

Lemma 11. *The following are equivalent: (a) $D(\Theta; L) \ge \lfloor g(L) \rfloor$ for all $L \in [1, L_{\max}]$; (b) $D(\Theta; L_k) \ge k$ for $k = 1, \dots, d/2$; (c) $\theta_{(k)} \le c_k$ for $k = 1, \dots, d/2$.*

Proof. (b) ⇔ (c): $D(\Theta; L) \ge k$ iff at least k frequencies are at most $1/L$, iff $\theta_{(k)} \le 1/L$; apply at $L = L_k$ with $1/L_k = c_k$. (a) ⇒ (b): $g(L_k) = k$ is an integer, so $\lfloor g(L_k) \rfloor = k$. (c) ⇒ (a): fix L and let $k := \lfloor g(L) \rfloor$; if $k = 0$ there is nothing to prove. Otherwise $k \le g(L)$ rearranges to $\ln L \le (1 - 2k/d)\ln L_{\max} = \ln L_k$, hence $L \le L_k$, hence $\theta_{(k)} \le c_k = 1/L_k \le 1/L$ and $D(\Theta; L) \ge k = \lfloor g(L) \rfloor$. □

Quantization is load-bearing. The real-valued requirement $D \ge g$ is false for the geometric comb itself: at $d = 4$ and $L_{\max} = 100$, the comb $\{1, 0.1\}$ has $g(3) = 1.523$ but $D(\Theta; 3) = 1$. Worse, the real-valued requirement changes the answer. Since D is integer-valued, $D \ge g$ is equivalent to $D \ge \lceil g \rceil$, and the argument of Lemma 11 then binds one rung lower: for L approaching L_{k-1} from below, $\lceil g(L) \rceil = k$ requires $\theta_{(k)} \le 1/L$ for every such L, hence $\theta_{(k)} \le c_{k-1}$ in the limit (with $c_0 := 1/L_{\max}$); conversely the shifted chain $\theta_{(k)} \le c_{k-1}$ for all k restores $D \ge \lceil g \rceil$ by the same rearrangement. Every frequency is then at most $c_{d/2-1} = L_{\max}^{-2/d} < 1$: the resolution endpoint $\theta = 1$ is forbidden, and the coordinatewise-maximal spectrum becomes the geometric comb shifted down one rung, so the optimal-design claim would name the wrong comb. The floor is therefore the correct notion of scale-free coverage for an integer-valued resource, and Theorem 5(iii) is true only under it.

D.4 Proof of Theorem 5(iii): optimality

Proof. The comb $\theta_j^* = L_{\max}^{-2(j-1)/d}$ has ascending order statistics $\theta_{(k)}^* = L_{\max}^{-(1-2k/d)} = c_k$, so it satisfies chain (c) of Lemma 11 with equality in every coordinate, and $c_k \le c_{d/2} = 1$ keeps it inside the domain; hence $\Theta^* \in \mathcal{F}$. For any $\Theta \in \mathcal{F}$, chain (c) gives $\theta_{(k)} \le c_k = \theta_{(k)}^*$ for every k, and any $\Theta \ne \Theta^*$ in $\mathcal{F}$ has $\theta_{(k)} < c_k$ at some k: Θ^* is the unique coordinatewise-maximal element. Uniqueness of the maximizer for any functional that increases when any order statistic is raised follows directly.

For the budget, the layer-cake decomposition $\lfloor g(L) \rfloor = \sum_{k=1}^{d/2} \mathbf{1}[g(L) \ge k] = \sum_{k=1}^{d/2} \mathbf{1}[L \le L_k]$ gives $\int_1^{L_{\max}} \lfloor g \rfloor \, d\ln L = \sum_{k=1}^{d/2} \ln L_k = \ln L_{\max} \sum_{k=1}^{d/2} \left(1 - \frac{2k}{d}\right) = \frac{d-2}{4} \ln L_{\max}$. Pointwise domination makes this a lower bound on $\int D \, d\ln L$ for every $\Theta \in \mathcal{F}$. For Θ^*, part (i) evaluates the expenditure as $\sum_j \ln(1/\theta_j^*) = \ln L_{\max} \sum_{j=1}^{d/2} \frac{2(j-1)}{d} = \frac{d-2}{4} \ln L_{\max}$ (each $\ln(1/\theta_j^*) \le \ln L_{\max}$, so the minimum in part (i) is inactive), matching the floor and establishing the stated equality. □

Efficiency. Equality in the budget bound of Theorem 5(i) requires every $\theta_m \le 1/L_{\max}$: the all-slow spectrum, with no resolution. The comb instead dominates g while spending the least coverage any dominating spectrum can spend, tight against the coverage floor, and simultaneously maximizes every resolution functional of Proposition 2. Coverage-efficiency and resolution-optimality are achieved by the same unique spectrum.

D.5 Alternative spectra at matched Fisher budget

Theorem 5 fixes one criterion, coverage and resolution. A natural alternative is the flat-profile Fisher information $\sum_m \theta_m^2$ (Proposition 2) for resolution, and certified alias-free range (Theorem 4) in place of coverage. We therefore hold Fisher information fixed and compare alias behavior.

Each alternative spectrum is rescaled to the deployed comb's own budget $\sum_m \theta_m^2$, matched to a relative error below 5×10^{-4} on every spectrum, and we scan for the first ϵ-alias of its r fastest bands at $\epsilon = 0.25$ out to 2×10^6, as in Appendix F. Four alternatives are used. *Linear* spaces θ arithmetically between the deployed spectrum's fastest and slowest bands, holding the range fixed and changing only the spacing law. *Uniform* places every band at one frequency, $\theta_m = \sqrt{2\,\text{budget}/d}$, a degenerate single-scale control. *Log-uniform* draws $d/2$ frequencies log-uniformly over the deployed range,

giving the comb's log-density without its regularity; we report an ensemble of 32 draws. *Max-min* maximizes a soft minimum of the flat-profile kernel $S(L)$ over 200 offsets to $L_{\max}$, subject to the same budget, and is the only designed competitor of the four.

The eight fastest bands separate the designs. On every one of the eight distinct deployed spectra, the comb's top-eight set returns no alias within 2×10^6, and so does every one of the 32 log-uniform draws, on all eight spectra: 256 draws, 256 with no return. The max-min spectra likewise return on none but one of their members. Neither structured alternative ever achieves it: *linear* aliases on all eight spectra, at 2,821 to 12,761 tokens, and *uniform* at 25 to 31. The property that separates them is log-density of bands across the range, not the exact geometric progression: a random log-uniform draw has it, an arithmetic ramp does not, and a single repeated clock does not. This is what the scale-free argument of Theorem 5 predicts, since invariance under rescaling the unit of distance forces coverage affine in $\ln L$, which is a condition on log-density.

The two fastest bands separate nothing. Table 7 reports the top-two alias distance. The comb beats the degenerate control on all eight spectra and the arithmetic ramp on six, but it is an unremarkable member of the log-uniform ensemble, falling between its 34th and 84th percentile with a median of 69. The reason is visible in the ensemble spread: within a single family the top-two distance runs from 7 to 496 tokens, far wider than any difference between the named schedules. A two-band alias distance is where one lattice point first enters the ϵ-ball (Appendix F), so it moves in jumps and cannot rank designs; only the $r = 8$ statistic, a status rather than a distance, does.

Table 7: Top-two alias distance at $\epsilon = 0.25$ under a matched Fisher budget, in tokens; smaller is worse. Log-uniform is an ensemble of 32 draws, reported as median and range, with percentile where the deployed comb falls within it. The three shared-spectrum pairs of Table 1 give identical rows and are listed once. On eight fastest bands, not shown, the comb, every log-uniform draw, and the max-min spectra return no alias within 2×10^6, while linear and uniform alias on every spectrum.

				log-uniform, 32 draws			
Model	comb	linear	uniform	median	min	max	comb pct.
Mistral-7B v0.1	138	201	25	113	8	327	53
Gemma-7B	88	113	25	21	7	465	59
SmolLM2	245	63	25	99	8	372	81
Llama-3-8B	239	226	28	107	7	453	81
Llama-3.1-8B	239	226	28	73	8	379	84
Mistral-7B v0.3	195	63	29	104	8	434	78
Yi-6B	88	69	30	109	8	382	34
Granite-3.1	113	94	31	141	8	496	44

E Structural Guarantees and Robustness

E.1 Proof of Proposition 3

Proof. The perturbed and unperturbed logits differ by $\langle \rho(t)\, q,\, \rho(j)\, \delta_j \rangle$. Cauchy–Schwarz and unitarity give $|\langle \rho(t)\, q,\, \rho(j)\, \delta_j \rangle| \le \|q\|\, \|\delta_j\|$, and $\|\rho(j)\, \delta_j\| = \|\delta_j\|$ makes the bound independent of both the write position and the read position. The total-variation clause is Proposition 1(iii) with ε the uniform logit bound. □

E.2 Deployment-Time Query Bound and Retrieval Preservation

Two corollaries of Proposition 3 make it operational before any query arrives: the first removes the dependence on future inputs, the second connects the logit bound to the retrieved entry.

Proof of Corollary 4. For an RMSNorm with gain γ, $\mathrm{RMSNorm}(x) = \gamma \odot x/\mathrm{rms}(x)$ and $\|x/\mathrm{rms}(x)\| = \sqrt{d_{\mathrm{model}}}$ for every x, so $\|\mathrm{RMSNorm}(x)\| \le \sqrt{d_{\mathrm{model}}}\, \|\gamma\|_\infty$. The head's query is $q_t = W_q^{(h)} \mathrm{RMSNorm}(x_t) + b_q^{(h)}$, and the triangle inequality with the operator norm gives the bound. Gemma's normalization applies $1 + \gamma$, so $\|\gamma\|_\infty$ is read as $\|1 + \gamma\|_\infty$ there; a model whose attention

scaling replaces $1/\sqrt{d}$ by a constant (Granite-3.1) applies that constant inside the score, not to q_t, so it does not enter the bound. □

Proof of Corollary 5. By Proposition 3 every score moves by at most ε, so $s'_J \geq s_J - \varepsilon > s_j + \varepsilon \geq s'_j$ for every $j \neq J$. □

Weight bound against measured query norms. Table 8 evaluates Corollary 4 on the ten checkpoints: the largest per-head bound in each model, the largest and the 99th-percentile query norm measured over three documents of 8192 tokens at all positions and heads, and the per-layer ratio of bound to measured maximum. The bound holds at every live layer of every model, as it must. It is conservative: the ratio ranges from 2.7 (Qwen2-7B, whose query bias reaches a per-head norm of 160 and dominates both sides) to 92 (SmolLM2), with per-model medians between 7.5 and 14.2. Substituting the weight bound for $|q|$ in Proposition 3 yields an input-free certificate, looser than the measured-norm certificate of Table 4 by that ratio: an order of magnitude at median. Where a norm statistic of the deployment distribution is acceptable, the measured p_{99} is the practical substitute.

Table 8: Deployment-time query bound of Corollary 4 against measured query norms. Bound is the largest per-head value of $\|W_q^{(h)}\|_{\mathrm{op}}\sqrt{d_{\mathrm{model}}}\|\gamma\|_\infty + \|b_q^{(h)}\|$ in the model; measured max and p_{99} are over all positions and heads of a three-document, 8192-token capture; ratio is bound over measured maximum per live layer, reported as minimum, median, and maximum over layers.

			measured $\|q\|$		ratio over layers		
Model	d_{model}	bound	max	p_{99}	min	median	max
Mistral-7B v0.1	4096	1,614	34	25	5.7	8.7	47
Gemma-7B	3072	1,754	109	61	5.8	9.8	21
SmolLM2	2048	3,399	37	27	10.0	12.9	92
Llama-3-8B	4096	830	52	36	4.7	7.6	24
Llama-3.1-8B	4096	754	45	33	5.0	7.5	21
Mistral-7B v0.3	4096	1,667	36	28	5.8	8.6	46
Qwen2-7B	3584	3,556	354	282	2.7	10.7	25
Qwen2.5-7B	3584	2,924	277	226	3.3	12.4	31
Yi-6B	4096	1,546	49	37	9.4	14.2	64
Granite-3.1	4096	10,071	205	125	8.4	12.8	59

E.3 Proof of Corollary 6

Storing each band's positional phase to finite precision induces a band-specific offset error w_m; part (ii) treats these jointly.

Proof. (i) Proposition 1(ii) with $|t_1 - t_2| \leq w$ bounds the perturbation by $\sum_m 2a_m \sup_{|\delta|\leq w} |\sin(\theta_m\delta/2)|$. On $\theta_m|\delta| \leq \pi$ the sine factor is increasing in $|\delta|$, and it never exceeds one, so the supremum is $\sin(\frac{1}{2}\min(\theta_m w, \pi))$.

(ii) Applying Proposition 1(ii) band by band with error w_m in band m, and $2|\sin(x/2)| \leq x$ for $x \geq 0$, gives the linear certificate $\sum_m 2a_m|\sin(\theta_m w_m/2)| \leq \sum_m \omega_m w_m$ with $\omega_m = a_m\theta_m$, valid for all $w_m \geq 0$.

For the allocation, a precision budget $\sum_m \ln(W_0/w_m) \leq B$ against a reference range W_0 is equivalent to $\prod_m w_m \geq P := W_0^{d/2}e^{-B}$. The certificate is increasing in each w_m, so the optimum lies on $\prod_m w_m = P$. By the AM–GM inequality over the $d/2$ bands, $\sum_m \omega_m w_m \geq \frac{d}{2}\big(\prod_m \omega_m w_m\big)^{2/d} = \frac{d}{2}\big(P\prod_m \omega_m\big)^{2/d}$, with equality iff all products $\omega_m w_m$ are equal. This forces $w_m^* = t/\omega_m$ with $t = (P\prod_j \omega_j)^{2/d}$, which satisfies the constraint with equality ($\prod_m w_m^* = t^{d/2}/\prod_m \omega_m = P$) and gives every band the same certified contribution $\omega_m w_m^* = t$. If some w_m^* exceeds its cap (the reference range, or π/θ_m when the exact sine form is required), clamp it and re-solve on the remaining bands in the standard water-filling manner. □

E.4 Proof of Theorem 6

Proof. Decomposing both arguments, $s(q,k,L) - s(q^{\perp},k^{\perp},L) = \langle \rho(t)\, q^{\perp},\, \rho(j)\, k^{\mathrm{pos}} \rangle + \langle \rho(t)\, q^{\mathrm{pos}},\, \rho(j)\, k \rangle$. Cauchy–Schwarz and unitarity bound the first term by $\|q^{\perp}\|\,\|k^{\mathrm{pos}}\| \le \|q\|\,\eta_k\|k\|$ and the second by $\eta_q\|q\|\,\|k\|$; summing gives the claim. The argument uses only Cauchy–Schwarz and unitarity; no compatibility between P_{pos} and ρ is invoked. □

Threshold shift in Theorem 3. Let $\varepsilon_\eta := (\eta_q + \eta_k)\, \sup_j \|q\|\,\|k_j\|$ denote the uniform logit perturbation. The signal decreases by at most ε_η and the interference maximum increases by at most ε_η, so the margin γ of Lemma 8 shifts by at most $2\varepsilon_\eta$, and both clauses of Theorem 3 hold with $S(L) \mp 2\varepsilon_\eta$ in place of $S(L)$. On the ensemble of Theorem 3, $\|q\| = \|k_J\|$ and the $\|k_j\|$ concentrate at one (Lemma 5), giving the normalized form $S(L) - C\eta$ for success and $S(L) + C\eta$ for failure, with C absorbing the concentration constants. The total-variation clause of the theorem is Proposition 1(iii) at $\varepsilon = \varepsilon_\eta$.

E.5 Proof of Corollary 7

Proof. The score is band-additive: $s(q,k,L) = \sum_m \mathrm{Re}\big(\bar{q}_m k_m e^{-i\theta_m L}\big)$. If P_{pos} selects the bands of $\mathcal{M}_{\mathrm{pos}}$, then $q_m^{\perp} = q_m$ and $k_m^{\perp} = k_m$ for every $m \notin \mathcal{M}_{\mathrm{pos}}$, so the contribution of every band outside $\mathcal{M}_{\mathrm{pos}}$ is identical for (q,k) and $(q^{\perp},k^{\perp})$. When $\mathcal{M}_{\mathrm{pos}}$ is disjoint from the coherent set $\{m : \theta_m L \le 1\}$, the coherent-band contribution at distance L is among the unchanged terms. That the band selectors are the only equivariant orthogonal projectors is the idempotent case of Theorem 2. □

F ADDITIONAL RESULTS

This appendix presents a thorough analysis of RoAM.

F.1 EVALUATION METHODOLOGY

Models. We evaluate ten base checkpoints. Table 9 lists parameters, depth, attention layout, and head dimension per model, with citations; spectrum-determined quantities are in Table 1. Pretrained Granite-3.1 consists of six effectively-zero k_proj layers ($20, 23, 24, 26, 27, 29$); all Granite-3.1 statistics in this paper cover the remaining 34 layers.

Table 9: Model roster: parameter class, layers, query and KV heads, and head dimension.

Model	Reference	Size	Layers	Heads (Q/KV)	d
Mistral-7B v0.1	Jiang et al. (2023)	7B	32	32 / 8	128
Gemma-7B	Team et al. (2024)	7B	28	16 / 16	256
SmolLM2	Allal et al. (2025)	1.7B	24	32 / 32	64
Llama-3-8B	Grattafiori et al. (2024)	8B	32	32 / 8	128
Llama-3.1-8B	Grattafiori et al. (2024)	8B	32	32 / 8	128
Mistral-7B v0.3	Jiang et al. (2023)	7B	32	32 / 8	128
Qwen2-7B	Yang et al. (2024)	7B	28	28 / 4	128
Qwen2.5-7B	Yang et al. (2025)	7B	28	28 / 4	128
Yi-6B	Young et al. (2024)	6B	32	32 / 4	128
Granite-3.1	Granite Team (2024)	8B	40	32 / 8	128

Tasks. We evaluate RoAM using the following.

- *RULER* (Hsieh et al., 2024): five subtasks varying the retrieval structure, 50 items each:
  - *single-key:* the single-needle form in RULER's needle-and-haystack format.
  - *multi-key:* 8 needles with distinct keys, one queried uniformly at random, so the remaining needles act as hard distractors.
  - *multi-value:* 4 values stored under one key; the response must produce all of them, scored with per-value partial credit.
  - *variable tracking:* a 4-hop chain of aliased variable assignments scattered through the text; the response must resolve the queried variable through the chain to its final value.
  - *ordering:* 4 needles at independent uniform depths; the response must list all stored values in order of first appearance, scored by the fraction of correctly ordered pairs, with omitted values counted as incorrectly ordered.
- *LongBench* (Bai et al., 2024): two retrieval subtasks, 50 items each:
  - *passage retrieval:* which of several candidate passages a summary was drawn from, adapted for base models with a completion-style answer scaffold and scored as an exact match on the answered passage number.
  - *HotpotQA:* multi-hop question answering over paired supporting documents, scored by token F1.
- *Zero-shot benchmarks*: 1000-item subsets, scored by length-normalized continuation log-likelihood, except where noted:
  - *LAMBADA* (Paperno et al., 2016): open-vocabulary prediction of a passage's final word, curated so the target requires the full context, scored by greedy multi-token exact match.
  - *HellaSwag* (Zellers et al., 2019): four-way multiple-choice sentence completion.
  - *PIQA* (Bisk et al., 2020): two-way choice between candidate solutions to a physical goal.
  - *ARC easy* and *challenge* (Clark et al., 2018): multiple-choice grade-school science questions with two difficulty splits.

Retrieval runs at 8192 target tokens for every model except Yi-6B, which runs at its declared 4096 maximum; no model is evaluated beyond its declared context.

Dataset and capture. PG-19 (Rae et al., 2020) is a long-form book corpus; we use ten test documents of 8192 tokens per seed (4096 for Yi-6B, its declared context) for perplexity and captures.

Captures record pre-rotation queries and keys at every layer and KV head, on the same documents, with queries strided by 8. Rotations are applied in analysis using each model's exact rotary buffer.

WikiText-103 (Merity et al., 2016) is used as a second corpus in Section F.12, and captures at 32768 and 65536 tokens in Section F.11.

F.2 Operational Equivalence

The KV cache, in its standard storage form and in the RoAM form, is run under a paired protocol: identical seeds, the same PG-19 documents, and retrieval and zero-shot task items, with three seeds per model. Lemma 1 predicts score-level identity; in finite precision, the two cache forms are distinct floating-point programs (rotation applied at write versus at read), so parity is an empirical question. The paired design holds evaluation items fixed across forms, while seeds resample the scored documents and task instances, which is the noise scale the main text reports. Table 10 reports per-task means and seed standard deviations for both cache forms on all ten models.

Table 10: Operational equivalence: mean ± sample standard deviation over three seeds. Identity denotes the standard post-rotation cache; RoAM the pre-rotation form. Blocks are perplexity, RULER, LongBench, and zero-shot. Column key: R:s, R:mk, R:mv, R:vt, R:ord are the RULER single-key, multi-key, multi-value, variable-tracking, and ordering subtasks; LB:pr, LB:hq are LongBench passage retrieval and HotpotQA; lam. = LAMBADA, hsw. = HellaSwag, piqa = PIQA, arc-e/c = ARC easy/challenge.

		Mistral-7B v0.1	Gemma-7B	SmolLM2	Llama-3-8B	Llama-3.1-8B	Mistral-7B v0.3	Qwen2-7B	Qwen2.5-7B	Yi-6B	Granite-3.1
PPL	identity	7.03 ± 0.44	9.82 ± 0.68	9.42 ± 0.86	8.46 ± 0.48	8.45 ± 0.48	6.99 ± 0.45	9.89 ± 0.84	9.38 ± 0.72	8.05 ± 1.81	6.20 ± 0.36
	RoAM	7.03 ± 0.44	9.82 ± 0.68	9.42 ± 0.86	8.46 ± 0.48	8.45 ± 0.47	6.99 ± 0.45	9.89 ± 0.84	9.38 ± 0.72	8.05 ± 1.81	6.20 ± 0.36
R:s	identity	0.44 ± 0.04	0.90 ± 0.02	0.93 ± 0.07	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00
	RoAM	0.44 ± 0.04	0.95 ± 0.01	0.92 ± 0.06	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00
R:mk	identity	0.39 ± 0.10	0.83 ± 0.05	0.39 ± 0.03	0.96 ± 0.05	0.98 ± 0.00	0.98 ± 0.02	1.00 ± 0.00	0.99 ± 0.02	0.44 ± 0.04	0.83 ± 0.07
	RoAM	0.39 ± 0.10	0.85 ± 0.11	0.39 ± 0.04	0.96 ± 0.05	0.98 ± 0.00	0.98 ± 0.02	1.00 ± 0.00	0.99 ± 0.02	0.45 ± 0.03	0.85 ± 0.05
R:mv	identity	0.25 ± 0.06	0.85 ± 0.04	0.26 ± 0.03	0.98 ± 0.01	1.00 ± 0.00	0.93 ± 0.03	0.95 ± 0.02	0.90 ± 0.03	0.52 ± 0.04	0.95 ± 0.04
	RoAM	0.25 ± 0.05	0.81 ± 0.01	0.26 ± 0.04	0.98 ± 0.01	1.00 ± 0.00	0.95 ± 0.01	0.95 ± 0.03	0.89 ± 0.03	0.53 ± 0.03	0.96 ± 0.03
R:vt	identity	0.43 ± 0.03	0.38 ± 0.05	0.17 ± 0.04	0.28 ± 0.06	0.56 ± 0.02	0.25 ± 0.06	0.06 ± 0.03	0.43 ± 0.10	0.46 ± 0.04	0.48 ± 0.10
	RoAM	0.43 ± 0.08	0.39 ± 0.08	0.15 ± 0.05	0.29 ± 0.04	0.55 ± 0.03	0.26 ± 0.05	0.05 ± 0.02	0.43 ± 0.12	0.47 ± 0.04	0.49 ± 0.13
R:ord	identity	0.12 ± 0.05	0.79 ± 0.05	0.10 ± 0.02	0.85 ± 0.04	0.87 ± 0.03	0.91 ± 0.01	0.87 ± 0.02	0.80 ± 0.02	0.25 ± 0.04	0.89 ± 0.04
	RoAM	0.12 ± 0.05	0.80 ± 0.03	0.13 ± 0.01	0.85 ± 0.03	0.87 ± 0.01	0.91 ± 0.01	0.87 ± 0.02	0.80 ± 0.02	0.24 ± 0.03	0.90 ± 0.05
LB:pr	identity	0.22 ± 0.04	0.25 ± 0.05	0.26 ± 0.05	0.46 ± 0.02	0.37 ± 0.01	0.34 ± 0.06	0.39 ± 0.06	0.34 ± 0.08	0.15 ± 0.02	0.37 ± 0.01
	RoAM	0.23 ± 0.03	0.25 ± 0.05	0.24 ± 0.07	0.46 ± 0.03	0.36 ± 0.00	0.35 ± 0.06	0.41 ± 0.04	0.32 ± 0.07	0.15 ± 0.02	0.39 ± 0.01
LB:hq	identity	0.32 ± 0.02	0.30 ± 0.04	0.23 ± 0.02	0.31 ± 0.02	0.33 ± 0.02	0.36 ± 0.02	0.55 ± 0.02	0.42 ± 0.04	0.38 ± 0.03	0.48 ± 0.01
	RoAM	0.34 ± 0.00	0.29 ± 0.04	0.24 ± 0.00	0.32 ± 0.01	0.33 ± 0.02	0.36 ± 0.02	0.55 ± 0.02	0.42 ± 0.03	0.36 ± 0.04	0.47 ± 0.02
lam.	identity	0.75 ± 0.02	0.67 ± 0.01	0.67 ± 0.01	0.76 ± 0.01	0.74 ± 0.01	0.76 ± 0.01	0.71 ± 0.01	0.72 ± 0.01	0.68 ± 0.02	0.76 ± 0.01
	RoAM	0.75 ± 0.02	0.67 ± 0.01	0.67 ± 0.01	0.76 ± 0.01	0.74 ± 0.01	0.76 ± 0.01	0.71 ± 0.01	0.72 ± 0.01	0.68 ± 0.02	0.76 ± 0.01
hsw.	identity	0.80 ± 0.01	0.79 ± 0.01	0.71 ± 0.02	0.79 ± 0.02	0.80 ± 0.01	0.80 ± 0.01	0.78 ± 0.02	0.77 ± 0.02	0.71 ± 0.01	0.80 ± 0.01
	RoAM	0.80 ± 0.01	0.79 ± 0.01	0.71 ± 0.02	0.79 ± 0.02	0.80 ± 0.01	0.80 ± 0.01	0.78 ± 0.02	0.77 ± 0.02	0.71 ± 0.01	0.80 ± 0.01
piqa	identity	0.81 ± 0.01	0.81 ± 0.01	0.78 ± 0.01	0.81 ± 0.00	0.81 ± 0.00	0.81 ± 0.01	0.80 ± 0.01	0.79 ± 0.01	0.78 ± 0.01	0.81 ± 0.00
	RoAM	0.81 ± 0.01	0.81 ± 0.01	0.78 ± 0.01	0.81 ± 0.00	0.81 ± 0.00	0.81 ± 0.01	0.80 ± 0.01	0.80 ± 0.00	0.78 ± 0.01	0.81 ± 0.01
arc-e	identity	0.65 ± 0.01	0.57 ± 0.01	0.66 ± 0.02	0.69 ± 0.01	0.65 ± 0.01	0.64 ± 0.01	0.71 ± 0.02	0.64 ± 0.01	0.61 ± 0.01	0.63 ± 0.01
	RoAM	0.65 ± 0.01	0.57 ± 0.00	0.66 ± 0.02	0.69 ± 0.01	0.65 ± 0.01	0.64 ± 0.01	0.71 ± 0.02	0.64 ± 0.01	0.61 ± 0.01	0.63 ± 0.01
arc-c	identity	0.43 ± 0.00	0.40 ± 0.00	0.41 ± 0.01	0.44 ± 0.00	0.41 ± 0.01	0.43 ± 0.00	0.46 ± 0.00	0.42 ± 0.00	0.39 ± 0.00	0.43 ± 0.01
	RoAM	0.43 ± 0.00	0.40 ± 0.00	0.41 ± 0.01	0.45 ± 0.00	0.41 ± 0.00	0.43 ± 0.00	0.46 ± 0.00	0.41 ± 0.00	0.40 ± 0.00	0.43 ± 0.01

Three patterns hold across the grid. Perplexity matches between the two forms to within 0.024% on every model, against an identity-form seed spread of 58%: the cache form matters less than the document sample by a factor of 2457×. The 120 task-model cells have a form difference less than the identity form's own seed standard deviation. Model capability, rather than the cache form, drives the spread: Mistral-7B v0.1 and SmolLM2 sit near 0.1 on ordering while the Llamas and Granite-3.1 clear 0.85, and both forms track that variation together. Yi-6B holds 1.00 on single-key while its harder subtasks sit lower (multi-key 0.44, ordering 0.25).

Cost of the RoAM frame. Lemma 1 leaves a choice of where the rotation is applied, and the two choices do not cost the same. Rotating the stored keys at read time is $O(Nd)$ per decoding

step; rotating the query by the negated offset is $O(d)$ and gives the same logits. Table 11 times all three paths on Llama-3.1-8B at batch size one, median of 100 decoding steps after 20 warmup, with the first 64 generated token ids identical across them. The query-side path tracks the standard post-rotation cache to within 1% at every cache length, while rotating the keys costs 1.4 to 88.2 milliseconds per step, growing by a factor of 63 as the cache grows by 64. Caching in RoAM form therefore costs nothing at inference when the rotation is read off the query.

Table 11: Milliseconds per decoding step on Llama-3.1-8B, median of 100 steps after 20 warmup, batch size one, bf16, synchronized. *post*: standard cache, keys stored rotated, query rotated at read time. *query*: keys stored pre-rotation, query rotated by the negated offset. *keys*: keys stored pre-rotation and rotated at read time. Interquartile spread is below 0.3% of the median in every cell.

Cache	1024	4096	16384	65536
post	16.86	18.56	24.49	51.94
query	16.91	18.61	24.55	52.45
keys	18.26	21.15	41.52	140.17

F.3 Per-Head Tuning Curves

Figure 5 plots the empirical distribution of per-head detrended correlations for every model against the permutation-null band and the cross-model median.

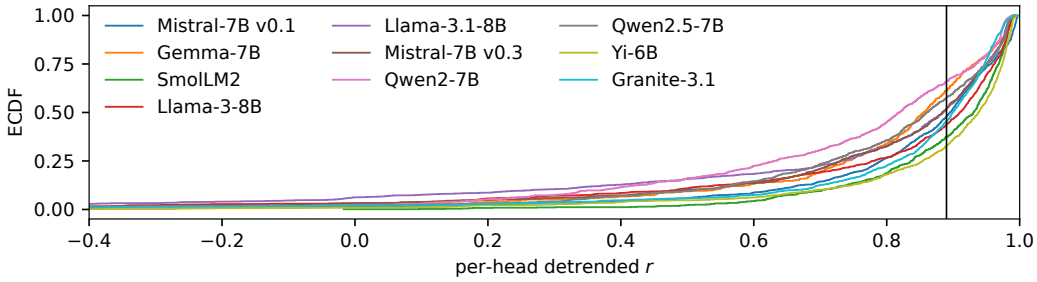


Figure 5: Empirical CDFs of per-head detrended correlation over live heads, per model.

Measured curves pool logits over all pairs at distance $L \geq 1$ into 24 logarithmic bins; the bin containing no integer distances is masked globally. Predictions follow Corollary 2, with per-band cross-covariances estimated over all pairs of queries and keys of the head (256 subsampled keys per document, ordered and unordered pairs alike) and both sides scaled by $1/\sqrt{d}$. Detrending subtracts the across-head mean of the measured and predicted curves of the model, computed over live heads, from each head's curves before correlating over unmasked bins. Controls comprise within-layer and global permutations of predictions across heads (5 rounds, reported raw and detrended), a leave-one-out shared-curve null, and a fixed-lag screen flagging heads whose nearest-bin logit exceeds the remaining curve by 3 standard deviations. Heads with constant curves are excluded from means and medians. Figure exemplars are the live, unflagged head whose seed-mean detrended correlation is closest to the model median, with curves averaged over three seeds.

The distributions are tight and well separated from their nulls. Per-model medians span 0.819 (Qwen2-7B) to 0.938 (Yi-6B), a range of 0.12 across six RoPE bases and three head dimensions, and the cross-model median of 0.890 sits within 0.005 of the pooled value. The within-layer null band (0.13 to 0.31) is the binding control since heads in one layer share both spectrum and depth; every model clears it by at least 0.62. The global permutation null is near zero ($|r| \leq 0.05$), confirming that the shared decay removed by detrending carried the resemblance no further.

Figure 6 plots the lowest-scoring live head per model, selected by seed-mean detrended correlation.

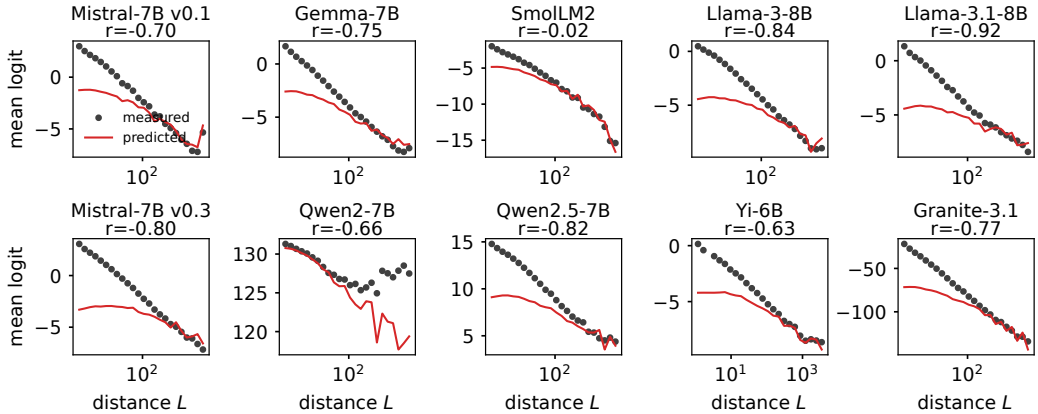


Figure 6: Lowest-scoring live head per model, selected by seed-mean detrended correlation (fixed-lag flag annotated where set); conventions as in Figure 2.

These heads are anticorrelated rather than uncorrelated: the worst head is strongly negative in nine of ten models, with a median of -0.76 and Llama-3.1-8B reaching -0.92; SmolLM2's worst head is the exception at -0.06, essentially uncorrelated. A prediction that tracks the measured curve with the sign reversed is a phase error, not an absent signal; thus, the failure mode is a mis-estimated band

phase rather than a head the spectrum fails to describe. The worst heads also sit deeper in the stack than the exemplars, at a median 52% of depth against 21%.

F.4 Positional Aliasing

For a band set, the operator deviation at offset Δ is $\max_m 2|\sin(\theta_m \Delta/2)|$, and the set is ϵ-aliased where this falls below $\epsilon = 0.25$. The deviation starts at zero, so we record two distances: the onset, where it first exceeds ϵ, and the alias distance, where it first returns below. Predictions were computed from model configurations alone and recorded in advance; measured scans use each model's captured rotary buffer. Table 12 reports both, and all 80 (onset, alias) pairs agree to the integer on all three seeds, confirming that configuration-reconstructed spectra reproduce the deployed buffers, including the one scaled spectrum.

Table 12: Aliasing scans, all band sets: entries are onset/alias distances under the exact operator criterion at $\epsilon = 0.25$; — denotes no return within 2×10^6. Registered (configuration-derived) and measured (buffer-derived) values agree on every entry; the table shows the common values.

Model	top-2	top-4	top-8	top-16	low-2	low-4	low-8	low-16
Mistral-7B v0.1	1/138	1/2438	1/—	1/—	1880/327942	1410/—	793/—	251/—
Gemma-7B	1/88	1/19547	1/—	1/—	2171/56138	1880/—	1410/—	793/—
SmolLM2	1/245	1/5837	1/—	1/—	15610/—	7478/—	1717/—	91/—
Llama-3-8B	1/239	1/11693	1/—	1/—	83168/—	55191/—	24304/—	4714/—
Llama-3.1-8B	1/239	1/11693	1/—	1/—	665342/—	441522/—	194432/—	37705/—
Mistral-7B v0.3	1/195	1/27508	1/—	1/—	162772/—	105701/—	44574/—	7927/—
Qwen2-7B	1/195	1/27508	1/—	1/—	162772/—	105701/—	44574/—	7927/—
Qwen2.5-7B	1/195	1/27508	1/—	1/—	162772/—	105701/—	44574/—	7927/—
Yi-6B	1/88	1/5749	1/—	1/—	773936/—	477928/—	182254/—	26504/—
Granite-3.1	1/113	1/12051	1/—	1/—	1514704/—	915330/—	334255/—	44574/—

Three observations follow. Onset for the top sets is $\Delta = 1$ everywhere, since the fastest frequency is $\theta_1 = 1$ in every deployed comb, and one step already exceeds ϵ. Widening a low set moves its onset earlier, monotonically in all ten models, because each added band is faster than the last. The top-two alias distance is short, 88 to 245 tokens, while at this tolerance no top-eight or top-sixteen set returns within 2×10^6: two bands beat against each other on a period set by their frequency gap, and adding bands makes simultaneous return exponentially harder, which is Theorem 4(ii) read across r.

Figure 7 plots the top-2 and top-4 deviation curves against distance, with the registered alias integers marked. Per head, 20% of 2864 heads alias within 2×10^6, and the aliasing heads are the sparse ones: median effective band count 4 against 11 for the rest. The rate varies widely by model, from 3.8% (Gemma-7B) to 48% (SmolLM2), tracking head sparsity rather than the base. Decision margins are finite: across returning entries, the flip tolerance exceeds the float32 error scale of the relevant bands by factors from 2.4 upward (tightest: Mistral-7B v0.1 lowest-2), so integer-exact reproduction requires scanning float32 buffers in float64 arithmetic, as both implementations do.

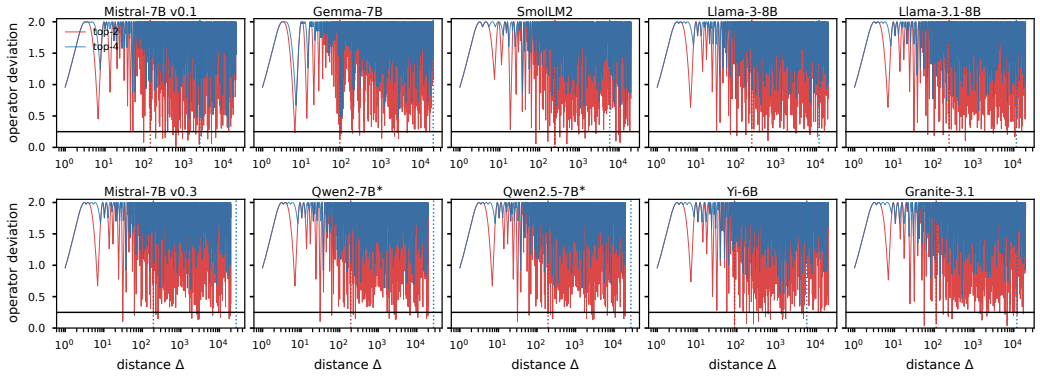


Figure 7: Operator deviation of each spectrum's top-2 and top-4 band sets against distance; the horizontal line is $\epsilon = 0.25$, and the dotted verticals mark the pre-registered alias integers, where the curves first return below threshold. *shared spectrum.

Sensitivity to the tolerance. The registry above fixes $\epsilon = 0.25$, a fraction of the logit scale: an ϵ-alias moves any score by at most $\epsilon\|q\|\|k\|$ and the softmax weights by at most $\beta\epsilon\|q\|\|k\|/2$ (Propo-

sition 1), so $\epsilon = 0.25$ admits returns that could shift a score by a quarter of its largest possible magnitude. Table 13 repeats the scan over $\epsilon \in \{0.1, 0.25, 0.5, 1\}$, to 2×10^6 as before.

What survives the change of tolerance is the ordering, not the distances. More retained bands buy a longer alias-free range at every ϵ, on every model. The top-sixteen set returns no alias at any tolerance we scan, and the top-eight set only once below $\epsilon = 1$, on Yi-6B at $1{,}557{,}872$, beyond any deployed context. The per-head rate rises with the tolerance, from 10.7% at $\epsilon = 0.1$ to 68.4% at $\epsilon = 1$, with the $\epsilon = 0.25$ row reproducing the 20% of the registry; the three models carrying the sparsest heads hold the three highest rates throughout.

The alias distance itself moves, and not smoothly. It is where a lattice point first falls inside the ϵ-ball, so it jumps as the ball admits new points. The top-two range runs from 88 to $1{,}552$ tokens at $\epsilon = 0.1$ and from 6 to 18 at $\epsilon = 1$. Band-sparse edits create returns at all four tolerances.

Table 13: Aliasing across the tolerance grid: first alias distance for the top-k band sets, as the range over the ten models, and the pooled fraction of the 2864 live heads whose own band support aliases. Scans to 2×10^6, spectrum only; — denotes no return within the scan. At $\epsilon = 0.1$ two models (SmolLM2, Yi-6B) also do not return on the top-four set and are excluded from that range; the top-eight entry at $\epsilon = 0.5$ is Yi-6B alone.

ϵ	top-2	top-4	top-8	top-16	heads aliasing
0.1	88 to 1,552	326,311 to 1,437,907	—	—	10.7%
0.25	88 to 245	2,438 to 27,508	—	—	20.3%
0.5	13 to 88	94 to 4,335	1,557,872	—	38.1%
1	6 to 18	7 to 208	377 to 32,119	—	68.4%

F.5 What band ablation costs outside retrieval

Table 14 reports zero-shot accuracy on the five tasks of Appendix F for the $f = 1/8$ and $f = 1/2$ cells of both arms, using the same three models and three seeds as Figure 3.

Table 14: Band ablation, three-seed means. Retrieval is the seven-task macro of Appendix F.1 and zero-shot the five-task macro; Δ columns are differences from each model's identity row.

Model	condition	PPL	retrieval	zero-shot	Δretr.	Δzero-shot
Gemma-7B	identity	9.82	0.613	0.646		
	slow, $f = 1/8$	5.4×10^6	0.000	0.458	−0.613	−0.188
	slow, $f = 1/2$	2.9×10^5	0.000	0.462	−0.613	−0.184
	fast, $f = 1/8$	9.92	0.602	0.643	−0.011	−0.003
	fast, $f = 1/2$	10.87	0.491	0.628	−0.122	−0.018
Llama-3.1-8B	identity	8.45	0.729	0.683		
	slow, $f = 1/8$	259.7	0.005	0.522	−0.724	−0.161
	slow, $f = 1/2$	1860.8	0.000	0.448	−0.729	−0.235
	fast, $f = 1/8$	8.66	0.717	0.677	−0.012	−0.006
	fast, $f = 1/2$	10.09	0.617	0.628	−0.111	−0.055
Granite-3.1	identity	6.20	0.717	0.686		
	slow, $f = 1/8$	413.7	0.000	0.520	−0.716	−0.166
	slow, $f = 1/2$	792.0	0.000	0.379	−0.717	−0.307
	fast, $f = 1/8$	6.50	0.671	0.675	−0.045	−0.011
	fast, $f = 1/2$	8.06	0.547	0.515	−0.169	−0.171

Retrieval degrades further than zero-shot accuracy in eleven of twelve cells, in both absolute and relative terms; the exception is Granite-3.1 on the fast arm at $f = 1/2$, where the two are within 0.002. The asymmetry is large on the slow arm: removing an eighth of the slow bands costs 99% to 100% of retrieval and 24% to 29% of zero-shot accuracy, and removing half costs 100% and 28% to 45%. It is also large on the fast arm at $f = 1/8$, where retrieval falls by 2% to 7% and zero-shot accuracy by 0.5% to 1.6%. It narrows at $f = 1/2$ on the fast arm, to 11% against 8% on Llama-3.1-8B, and closes on Granite-3.1, where both metrics lose 25%.

Short-range grammatical discrimination. Removing the fast bands leaves retrieval and perplexity nearly intact, which raises the question of whether they carry a short-range function the retrieval suite cannot see. We score BLiMP (Warstadt et al., 2020), 67 paradigms of minimal pairs, on the same cells, with each pair placed at an offset of 4096 tokens behind PG-19 filler so that the contrast stays local while the cache is long. Table 15 reports the macro. Neither fast arm moves it consistently: the sign of the shift differs across models, and within a model the paradigms move in both directions, 16% to 75% of them downward. Removing an eighth of the slow bands moves the macro by −0.136 to −0.181 with 87% to 91% of paradigms down on every model, so the probe registers damage where damage exists. That damage is not specific to grammar: on the same cells zero-shot accuracy falls by 24% to 29% against BLiMP's 19% to 26%. The fast half carries no grammatical discrimination this probe can detect.

Table 15: BLiMP macro under band ablation, three-seed means over 67 paradigms of 500 minimal pairs each, scored at an offset of 4096 tokens. Δ columns are differences from each model's identity row. At zero filler Llama-3.1-8B scores 0.785 against 0.729 here, so the filler itself costs 0.056; every arm shares the same filler within a seed.

Model	identity	Δ fast, $f = 1/8$	Δ fast, $f = 1/2$	Δ slow, $f = 1/8$
Gemma-7B	0.709	+0.011	+0.034	−0.181
Llama-3.1-8B	0.729	−0.006	+0.001	−0.136
Granite-3.1	0.732	+0.014	−0.024	−0.157

F.6 RANDOM-BAND CONTROL

In Section 7, we show that removing the slow half damages retrieval and removing the fast half does not, but both remove the same number of coordinates from opposite ends of an ordered list. We repeat the ablation with the removed set drawn uniformly at random from the spectrum, at the same band count, three draws per model per fraction. The draw is reseeded with the run, so the three seeds differ in which bands are removed as well as in the scored documents. Because the draw is uniform, each lands at a different slow-band count, and the compressor logs which bands it took. Table 16 orders the $f = 1/8$ runs by that count, with slow-end and fast-end removal as the endpoints.

Table 16: Random-band control at $f = 1/8$, ordered by the number of removed bands that lie in the slow half. Removed is the band count, identical across conditions within a model; slow is how many of those bands were slow. The fast and slow rows are the deterministic removals of Figure 3; the random rows are one draw per seed.

Model	removed from	removed	slow	R:ord	PPL
Gemma-7B	fast end	16	0	0.798	9.92
	random (seed 123)	16	6	0.000	22.6
	random (seed 456)	16	7	0.000	318.4
	random (seed 42)	16	9	0.030	15.1
	slow end	16	16	0.000	5.4×10^{6}
Llama-3.1-8B	fast end	8	0	0.850	8.66
	random (seed 123)	8	3	0.787	9.9
	random (seed 42)	8	5	0.000	15.8
	random (seed 456)	8	5	0.307	12.9
	slow end	8	8	0.000	259.7
Granite-3.1	fast end	8	0	0.773	6.50
	random (seed 123)	8	3	0.333	7.3
	random (seed 42)	8	5	0.000	11.5
	random (seed 456)	8	5	0.130	7.8
	slow end	8	8	0.000	413.7

Retrieval falls monotonically in the number of slow bands removed while the total removed is held fixed. On Llama-3.1-8B and Granite-3.1, ordering accuracy runs 0.81 at zero slow bands, 0.56 at three, 0.11 at five, and 0.00 at eight, with Spearman -0.92 between slow-band count and accuracy on both models. Removing an eighth of the spectrum does damage retrieval, and how much is set by how much of that eighth was slow. Gemma-7B shows no gradient. Every draw took six or more slow bands and collapsed, which is consistent with it being the most fragile model in the set, since $d = 256$ gives it 128 bands and six slow bands out of 64 already suffice. At $f = 1/2$ the control carries no information, as a random half contains 14 to 17 slow bands against the eight that already destroy retrieval; all three models collapse there.

This random-band control and the matched-norm experiment of Table 2 together isolate placement. The first holds the number of removed coordinates fixed while varying where they fall; the second holds the perturbation norm fixed and removes no coordinate at all.

F.7 Capacity Frontier

We realize the retrieval event of Theorem 3 on each deployed spectrum. Each trial draws N iid Gaussian keys ($\mathcal{N}(0, 1/d)$ per coordinate), places a content-echo query at offset L with distractor offsets uniform in $[1, 4L)$, and scores success as argmax over the store. The grid is $L \in \{10^3, 5 \times 10^3, 2 \times 10^4, 10^5\}$ by $N \in \{4, 16, 64, 256, 1024\}$, with 200 trials per cell and three seeds. For a content-echo pair whose key coordinates are iid $\mathcal{N}(0, 1/d)$, band m contributes $(a_m^2 + b_m^2)\cos(\theta_m L)$ with $a_m, b_m \sim \mathcal{N}(0, 1/d)$, so each band's weight is a scaled χ_2^2 with mean $2/d$ and variance $4/d^2$. The realized score therefore has mean $S(L)$ and spread

$$S(L) = \frac{2}{d}\sum_{m=1}^{d/2}\cos(\theta_m L), \qquad \sigma(L) = \sqrt{\frac{4}{d^2}\sum_{m=1}^{d/2}\cos^2(\theta_m L)}\,.$$

Both are determined by the spectrum and d alone: σ is the finite-d fluctuation of the signal about its mean, not an empirical error bar, and it sets the width of the unclassified band. A cell is predicted to be a success when $S(L)$ exceeds the threshold $\sqrt{2\ln(N-1)/d}$ by σ, a failure when it falls short by σ, and is left unclassified between the two.

The classification is exact on the decided cells. All 63 predicted successes attain a rate of ≥ 0.9, and all 74 predicted failures rate ≤ 0.5, with no violations in either direction; the remaining 63 cells fall in the unclassified band. Every in-grid crossing of rate 0.5 lands inside the store-size bracket $N = \exp(d\,(S(L) \mp \sigma)^2/2)$ that the two thresholds predict: 11 crossings on all three seeds and 2 on a subset, marked b in Table 17 (Figure 8).

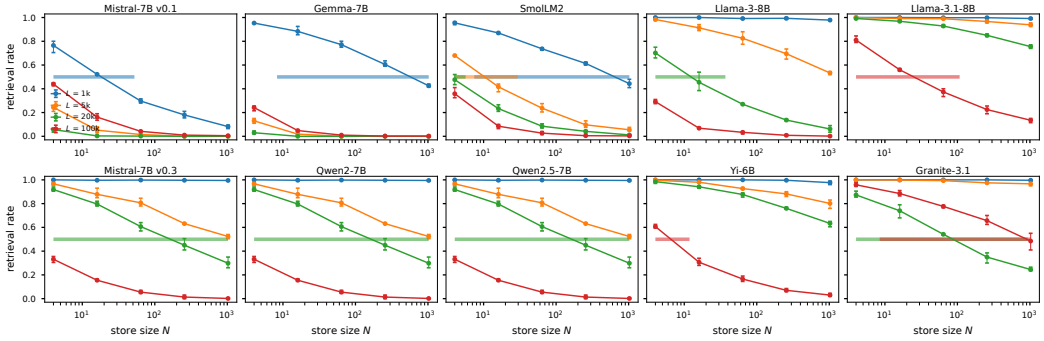


Figure 8: Capacity frontier on deployed spectra: retrieval rate against store size N at four offsets L, seed means with min-max whiskers over three seeds. Horizontal brackets at rate 0.5 mark each offset's threshold-predicted crossover interval; every empirical crossing falls inside its bracket.

Table 17 reports the full crossover data: per spectrum and offset, the empirical store size at rate 0.5 against the threshold-predicted interval.

Table 17: Capacity crossovers, per model and offset: $S(L)$ of the deployed spectrum, the threshold-predicted crossover bracket in store size, and the empirical rate-0.5 crossing (log-interpolated; mean $\pm$ sample standard deviation over three seeds). Entries > 1024 or < 4 lie outside the simulated grid; every in-grid crossing falls inside its predicted bracket.

Model	L	$S(L)$	predicted $[N_{\mathrm{lo}}, N_{\mathrm{hi}}]$	empirical N^*	in
Mistral-7B v0.1	1000	$+0.159$	$[1.36,\ 52.37]$	18 ± 1	✓
Mistral-7B v0.1	5000	-0.007	$[1.00,\ 1.51]$	< 4	—
Mistral-7B v0.1	20000	-0.106	$[1.00,\ 1.01]$	< 4	—
Mistral-7B v0.1	100000	$+0.056$	$[1.00,\ 3.57]$	< 4	—
Gemma-7B	1000	$+0.193$	$[8.29,\ 4548]$	574 ± 46	✓
Gemma-7B	5000	-0.029	$[1.00,\ 1.14]$	< 4	—
Gemma-7B	20000	-0.087	$[1.00,\ 1.07]$	< 4	—
Gemma-7B	100000	-0.002	$[1.00,\ 1.56]$	< 4	—
SmolLM2	1000	$+0.390$	$[7.43,\ 7679]$	673 ± 152	✓
SmolLM2	5000	$+0.190$	$[1.10,\ 29.61]$	10 ± 1	✓
SmolLM2	20000	$+0.098$	$[1.00,\ 5.65]$	4 [b]	✓
SmolLM2	100000	$+0.043$	$[1.00,\ 2.48]$	< 4	—
Llama-3-8B	1000	$+0.492$	$[1.4 \times 10^4,\ 8.7 \times 10^9]$	> 1024	—
Llama-3-8B	5000	$+0.293$	$[11.91,\ 1.6 \times 10^4]$	> 1024	—
Llama-3-8B	20000	$+0.143$	$[1.16,\ 36.94]$	14 ± 5	✓
Llama-3-8B	100000	$+0.001$	$[1.00,\ 1.83]$	< 4	—
Llama-3.1-8B	1000	$+0.560$	$[4.4 \times 10^5,\ 2.8 \times 10^{12}]$	> 1024	—
Llama-3.1-8B	5000	$+0.435$	$[1168,\ 1.1 \times 10^8]$	> 1024	—
Llama-3.1-8B	20000	$+0.360$	$[78.79,\ 7.1 \times 10^5]$	> 1024	—
Llama-3.1-8B	100000	$+0.170$	$[1.37,\ 107]$	25 ± 2	✓
Mistral-7B v0.3	1000	$+0.491$	$[1.6 \times 10^4,\ 5.9 \times 10^9]$	> 1024	—
Mistral-7B v0.3	5000	$+0.286$	$[8.95,\ 1.5 \times 10^4]$	> 1024	—
Mistral-7B v0.3	20000	$+0.235$	$[3.36,\ 1196]$	179 ± 77	✓
Mistral-7B v0.3	100000	$+0.036$	$[1.00,\ 2.82]$	< 4	—
Qwen2-7B	1000	$+0.491$	$[1.6 \times 10^4,\ 5.9 \times 10^9]$	> 1024	—
Qwen2-7B	5000	$+0.286$	$[8.95,\ 1.5 \times 10^4]$	> 1024	—
Qwen2-7B	20000	$+0.235$	$[3.36,\ 1196]$	179 ± 77	✓
Qwen2-7B	100000	$+0.036$	$[1.00,\ 2.82]$	< 4	—
Qwen2.5-7B	1000	$+0.491$	$[1.6 \times 10^4,\ 5.9 \times 10^9]$	> 1024	—
Qwen2.5-7B	5000	$+0.286$	$[8.95,\ 1.5 \times 10^4]$	> 1024	—
Qwen2.5-7B	20000	$+0.235$	$[3.36,\ 1196]$	179 ± 77	✓
Qwen2.5-7B	100000	$+0.036$	$[1.00,\ 2.82]$	< 4	—
Yi-6B	1000	$+0.491$	$[1.4 \times 10^4,\ 6.9 \times 10^9]$	> 1024	—
Yi-6B	5000	$+0.368$	$[86.45,\ 1.5 \times 10^6]$	> 1024	—
Yi-6B	20000	$+0.319$	$[23.69,\ 6.7 \times 10^4]$	> 1024	—
Yi-6B	100000	$+0.103$	$[1.01,\ 11.82]$	7 ± 0	✓
Granite-3.1	1000	$+0.584$	$[2.6 \times 10^6,\ 1.4 \times 10^{13}]$	> 1024	—
Granite-3.1	5000	$+0.453$	$[2243,\ 4.9 \times 10^8]$	> 1024	—
Granite-3.1	20000	$+0.221$	$[2.36,\ 931]$	88 ± 8	✓
Granite-3.1	100000	$+0.284$	$[8.47,\ 1.4 \times 10^4]$	818 [b]	✓

[b] crossing in-grid on a subset of seeds; the remainder stay above 0.5 at $N = 1024$ (Granite-3.1) or below at $N = 4$ (SmolLM2).

Three observations follow. The three shared-spectrum models agree to every digit, as they must, and the scaled Llama-3.1-8B spectrum holds $S(L) = +0.170$ at $L = 10^5$ where its unscaled counterpart has fallen to $+0.001$. $S(L)$ is not monotone in L: Mistral-7B v0.1 and Granite-3.1 both recover at 10^5 after a trough at 2×10^4, the almost-periodic behavior of Section 3, and the simulated rates track the recovery. The bracket narrows as $S(L)$ approaches the threshold; thus, the prediction is sharpest exactly where the crossing is observable: Granite-3.1 spans seven orders of magnitude at $L = 10^3$, where $S(L) = +0.584$, and $[2.36,\ 931]$ at 2×10^4, where it is $+0.221$.

F.8 Positional Leakage

The tiled probe concatenates eight copies of one 256-token block, so every token recurs at eight positions with identical content. For each such token, we measure how its key varies across occurrences, dropping the first copy, whose left context differs the most. Write $k_{j,h}^{(t)}$ for the pre-action key of head h at within-block offset j in tile t, with $t = 2, \ldots, 8$. The content estimate at that offset is the across-tile mean $\bar{k}_{j,h} = \frac{1}{7}\sum_t k_{j,h}^{(t)}$, and the residual $r_{j,h}^{(t)} = k_{j,h}^{(t)} - \bar{k}_{j,h}$ is the part that varies with position at fixed content. Then

$$\eta_k(h) \;=\; \left(\frac{1}{d} \sum_{c=1}^{d} \frac{\langle (r_{j,h,c}^{(t)})^2 \rangle_{t,j}}{\langle (k_{j,h,c}^{(t)})^2 \rangle_{t,j}} \right)^{1/2},$$

a dimensionless amplitude fraction in $[0, 1]$: 0 when a head's keys are identical across occurrences, 1 when content explains none of their energy. Because content is matched, this variation can arise only from position or residual context drift, so η_k is a conservative upper bound on position dependence; a positional component constant across positions is invisible to it by construction. Per-dimension injections at $\epsilon = 0.05$ and 0.2 are recovered to within 0.02 of the independent analytic value $\sqrt{\eta_0^2 + \epsilon^2}$, so the estimator moves at the expected magnitude on a planted signal of known size.

Figure 9 plots the bound against depth, and Table 18 summarizes it per model.

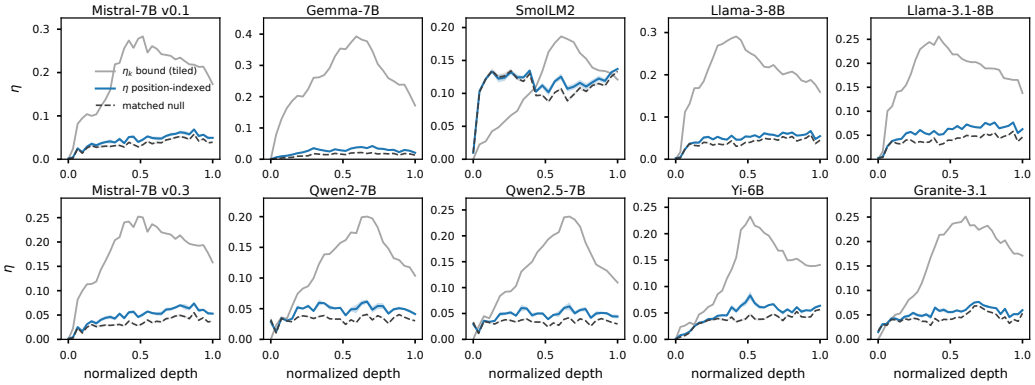


Figure 9: Positional leakage per model: the tiled upper bound η_k against normalized depth, three-seed means; the position-indexed and matched-null series are the subspace probe of Section F.12, plotted here for comparison. The bound is zero at the embedding, peaks mid-stack, and decays toward the readout; at every peak it is larger in the slow half of the spectrum than the fast half.

The profiles share a similar shape. η_k is exactly zero at layer 0 in all ten models, as it must be for an embedding that has seen no context, rises to a peak between 41% and 64% of depth, then falls back by a factor of 1.5 to 2.3 by the final layer. Leakage that survived the embedding would peak early, and leakage induced near the readout would peak late; a mid-stack maximum that decays afterward matches neither, and matches drift accumulated and then partly discarded through the residual stream. The bound itself is loose: at 0.19 to 0.39 at the worst layer, it admits a position-carrying subspace holding up to a sixth of the key energy and rules out anything larger.

The bound is not band-uniform. At each model's peak layer, the slow half carries 1.08 to 1.44 times the leakage of the fast half and is larger in all ten models. Corollary 7 shows that leakage outside the coherent set leaves retrieval at the corresponding distances unchanged; the measured concentration is the unfavorable case since the slow bands are the coherent ones at long range. The effect is small in absolute terms, and at this bound, the robustness shift of Theorem 6 reverses no capacity-frontier prediction, but the direction is the one that costs.

Table 18: Key-side position bound η_k, a per-layer mean over tiles averaged across three seeds. Median is over layers; peak is the maximum, with its layer index and fraction of depth. Slow/fast is the ratio of mean band leakage in the slow half of the spectrum to the fast half, at the peak layer.

Model	median	peak	layer	depth	slow/fast
Mistral-7B v0.1	0.22	0.28	16	50%	1.25
Gemma-7B	0.26	0.39	16	57%	1.08
SmolLM2	0.13	0.19	14	58%	1.21
Llama-3-8B	0.21	0.29	13	41%	1.10
Llama-3.1-8B	0.19	0.26	13	41%	1.15
Mistral-7B v0.3	0.20	0.25	15	47%	1.25
Qwen2-7B	0.13	0.20	18	64%	1.23
Qwen2.5-7B	0.14	0.24	18	64%	1.19
Yi-6B	0.14	0.23	16	50%	1.44
Granite-3.1	0.17	0.25	21	53%	1.23

F.9 CERTIFIED COMPRESSION

We snapshot each model's cache after an 8192-token prefill, run one compress and decompress cycle through each method's own implementation, and score sampled (query, key, offset) triples in the post-rotation frame; the scope is the key-side logit channel. Table 19 reports the per-model quality components behind the main-text aggregates, and Table 20 the asymmetry $\Delta\text{Retr} - \Delta\text{ZS}$ per cell. Quality deltas are computed seed by seed against the identity cache of the same seed, since the scored documents are resampled per seed.

Table 19: Per-model quality under each deployed method: ΔPPL (relative), ΔR (seven-task retrieval macro), and ΔZ (five-task zero-shot macro), three-seed means against the identity cache of the same seed. The Table 4 medians are computed from these values before rounding.

	KIVI-2			KVQuant-3			PALU			H2O		
Model	ΔPPL	ΔR	ΔZ	ΔPPL	ΔR	ΔZ	ΔPPL	ΔR	ΔZ	ΔPPL	ΔR	ΔZ
Mistral-7B v0.1	+2.7%	−0.02	−0.00	+0.6%	−0.01	−0.00	+9.2%	−0.06	−0.02	+16.7%	−0.27	−0.00
Gemma-7B	+124%	−0.51	−0.00	+1.8%	−0.02	−0.00	+76.2%	−0.33	−0.01	+11.3%	−0.53	−0.00
SmolLM2	+13.5%	−0.21	+0.00	+1.6%	−0.07	−0.00	$> 10^4$%	−0.33	−0.09	+12.4%	−0.28	+0.00
Llama-3-8B	+6.4%	−0.14	−0.00	+1.1%	−0.04	−0.00	+215%	−0.62	−0.03	+10.2%	−0.62	−0.00
Llama-3.1-8B	+5.9%	−0.08	+0.00	+1.1%	−0.01	−0.00	+46.8%	−0.19	−0.03	+11.1%	−0.65	+0.00
Mistral-7B v0.3	+3.2%	−0.11	+0.00	+0.6%	−0.02	+0.00	+9.0%	−0.08	−0.02	+9.2%	−0.61	+0.00
Qwen2-7B	+2.9%	−0.08	+0.00	+0.7%	−0.01	−0.00	+325%	−0.50	−0.09	+9.8%	−0.59	+0.00
Qwen2.5-7B	+3.3%	−0.08	−0.00	+0.7%	−0.02	−0.00	+656%	−0.61	−0.08	+9.5%	−0.62	−0.00
Yi-6B	+2.7%	−0.08	+0.00	+0.7%	−0.04	+0.00	+10.4%	−0.05	−0.01	+6.0%	−0.36	+0.00
Granite-3.1	+4.2%	−0.08	−0.00	+0.9%	−0.00	−0.00	+28.9%	−0.20	−0.02	+8.8%	−0.61	−0.00

Table 20: Degradation asymmetry between retrieval and zero-shot per cell, mean ± sample standard deviation over three seeds against the identity cache of the same seed. Negative means retrieval degrades more than zero-shot.

Model	KIVI-2	KVQuant-3	PALU	H2O
Mistral-7B v0.1	-0.021 ± 0.030	-0.004 ± 0.014	-0.045 ± 0.015	-0.268 ± 0.037
Gemma-7B	-0.505 ± 0.012	-0.023 ± 0.014	-0.319 ± 0.041	-0.532 ± 0.005
SmolLM2	-0.207 ± 0.013	-0.072 ± 0.015	-0.243 ± 0.016	-0.277 ± 0.004
Llama-3-8B	-0.139 ± 0.019	-0.039 ± 0.021	-0.585 ± 0.022	-0.617 ± 0.013
Llama-3.1-8B	-0.077 ± 0.011	-0.007 ± 0.006	-0.165 ± 0.026	-0.651 ± 0.003
Mistral-7B v0.3	-0.113 ± 0.032	-0.025 ± 0.006	-0.058 ± 0.029	-0.607 ± 0.014
Qwen2-7B	-0.076 ± 0.036	-0.010 ± 0.023	-0.416 ± 0.020	-0.592 ± 0.014
Qwen2.5-7B	-0.084 ± 0.022	-0.016 ± 0.005	-0.536 ± 0.037	-0.616 ± 0.026
Yi-6B	-0.076 ± 0.020	-0.036 ± 0.020	-0.042 ± 0.003	-0.361 ± 0.013
Granite-3.1	-0.078 ± 0.031	$+0.000 \pm 0.023$	-0.185 ± 0.045	-0.612 ± 0.040

Each reconstruction method contributes 38,292,416 triples over its thirty cells, 115 million in total, with zero violations; the identity control and the RoAM form contribute 38.3 million triples apiece, also without violation. H2O is verified per retained slot rather than per triple because its certificate is vacuous wherever it evicts: an evicted slot carries $\delta = -k$, so the bound degenerates to $\|q\|\|k\|$. At the audited ratio, it evicts 94% slots; every retained slot is exact, and at evicted slots, the ratio of measured shift to bound has a median 0.24 and maximum 0.57: the bound holds and carries no information there. Eviction is neither spectrally nor positionally neutral. 69% of the energy it removes lies in the slow half of the spectrum, and it evicts 96% to 99% of the first nine position deciles against 61% of the last, so it removes precisely the distant, slow-band structure that Figure 3 shows retrieval depends on.

Rank informativeness varies with error structure: per-model ρ spans $[0.18, 0.52]$ under KIVI-2, $[0.23, 0.55]$ under KVQuant-3, and $[0.29, 0.71]$ under PALU. Near-uniform quantization error carries little per-key rank information; structured low-rank error carries the most. PALU is also the method whose damage varies most across models, from 9.0% perplexity on Mistral-7B v0.3 to divergence on SmolLM2, whose $d = 64$ leaves the least room for a rank cut. Spectrum alone does not fix the outcome: under PALU the shared-spectrum pair Qwen2-7B and Qwen2.5-7B lands at asymmetries -0.416 and -0.536, and the scaled and unscaled Llama pair differs by a factor of 3.5.

Margin screen. Corollary 5 turns the logit bound into a per-read retrieval guarantee: the top entry survives any edit for which $\max_j \|q\|\|\delta_j\|$ is below half the read's margin. We evaluate the condition on every read of two documents of 4096 tokens, at every layer and head, for the three reconstruction methods in their own implementations (3.5 million reads over the ten base models), and record beside it the empirical fraction of reads whose top entry did survive. Table 21 restricts to reads whose top entry is at distance ≥ 64, the regime the paper is about; the unrestricted figures differ by at most 0.02 for the certified fraction.

Table 21: Margin screen on reads whose top entry is at distance ≥ 64: fraction certified preserved by Corollary 5 (cert.) and fraction whose top entry survived the edit (emp.), per method; for H2O, the median and 90th percentile of attention mass on evicted slots per read. Reads from two documents of 4096 tokens at every layer and head, methods in their own implementations; H2O keeps 512 slots.

	KIVI-2		KVQuant-3		PALU		H2O mass	
Model	cert.	emp.	cert.	emp.	cert.	emp.	p_{50}	p_{90}
Mistral-7B v0.1	0.005	0.731	0.037	0.898	0.060	0.876	0.52	0.87
Gemma-7B	0.000	0.927	0.132	0.985	0.002	0.892	0.45	0.87
SmolLM2	0.000	0.767	0.008	0.944	0.000	0.880	0.64	0.98
Llama-3-8B	0.000	0.947	0.024	0.979	0.025	0.970	0.29	0.78
Llama-3.1-8B	0.000	0.943	0.020	0.976	0.025	0.933	0.30	0.79
Mistral-7B v0.3	0.004	0.818	0.036	0.939	0.060	0.903	0.31	0.82
Qwen2-7B	0.000	0.868	0.001	0.933	0.000	0.859	0.59	0.97
Qwen2.5-7B	0.000	0.869	0.001	0.934	0.000	0.850	0.60	0.97
Yi-6B	0.001	0.929	0.022	0.966	0.035	0.957	0.36	0.91
Granite-3.1	0.000	0.726	0.002	0.833	0.001	0.735	0.00	1.00
median	0.000	0.868	0.021	0.941	0.013	0.886	0.41	0.89

The certified fraction is small at deployed compression rates: zero to 0.5% on KIVI-2, a median of 2.1% and at most 13% (Gemma-7B) on KVQuant-3, and a median of 1.3% and at most 6.0% on PALU, while the top entry survives empirically on 73% to 99% of reads. The gap is the price of a worst-case, distribution-free bound: the certificate takes the largest per-key error in the cache and charges it against every read, whereas the measured shift on the median read is below $\|q\|\|\delta_j\|$ by a factor of 6 to 25 (the tightness of Table 4). Read as a design target: a certifiably retrieval-preserving edit needs per-key error roughly an order of magnitude below what 2-bit and 3-bit quantization produce. H2O has no per-key bound; the corresponding statistic is the attention mass a read places on evicted slots, whose per-read mean is 0.29 to 0.58 across models at 87.5% of slots evicted, with a median of zero on Granite-3.1 and a 90th percentile at or above 0.78 on every model.

F.10 Capacity on Real Keys

Table 22 reports the capacity frontier of Figure 8 rerun on captured keys. For each model, we sample twelve (layer, head) pairs, draw stores of N keys without replacement from that head's captured pre-rotation pool, place the content-echo query at offset L with distractor offsets uniform in $[1, 4L)$, and score success as argmax over the store. The grid matches the synthetic one, $L \in \{10^3, 5{\times}10^3, 2{\times}10^4, 10^5\}$ by $N \in \{4, 16, 64, 256, 1024\}$, with 100 trials per cell and three seeds, giving 6960 cells. Predictions come from the head's own measured band energy: $S(L) = \sum_m e_m \cos(\theta_m L)$ with e_m, the normalized mean band energy of the pool. We score the margin $S(L) - \sqrt{2\ln(N-1)/d}$ against the measured rate.

Table 22: Capacity on captured keys: twelve (layer, head) pairs per model, 100 trials per cell, three seeds. ρ is Spearman between the predicted margin and the measured rate over all cells of that model; the three columns beside it hold one factor fixed and vary another, and slow is the median share of band energy in the slow half over the sampled heads. The rate columns are pooled means at the smallest and largest store size. Granite-3.1 contributes eight heads.

			ρ within one factor				rate	
Model	heads	ρ	vary N	vary L	vary head	slow	$N{=}4$	$N{=}1024$
Mistral-7B v0.1	12	$+0.69$	0.87	0.80	$+0.55$	0.74	0.39	0.01
Gemma-7B	12	$+0.70$	0.86	0.88	$+0.71$	0.64	0.38	0.03
SmolLM2	12	$+0.62$	0.90	0.78	$+0.49$	0.76	0.46	0.00
Llama-3-8B	12	$+0.82$	0.92	0.89	$+0.69$	0.64	0.56	0.04
Llama-3.1-8B	12	$+0.81$	0.95	0.86	$+0.67$	0.66	0.72	0.14
Mistral-7B v0.3	12	$+0.65$	0.94	0.81	$+0.42$	0.71	0.62	0.09
Qwen2-7B	12	-0.17	0.95	0.79	-0.68	0.87	0.69	0.21
Qwen2.5-7B	12	-0.16	0.96	0.77	-0.66	0.88	0.69	0.21
Yi-6B	12	$+0.68$	0.96	0.72	$+0.33$	0.69	0.63	0.06
Granite-3.1	8	$+0.73$	0.97	0.74	$+0.29$	0.78	0.64	0.03

Retrieval rate declines monotonically with store size in 1318 of 1392 (head, offset) curves, and pooled across models, the mean rate falls from 0.58 at $N = 4$ to 0.08 at $N = 1024$. The margin orders the measured rate at Spearman $+0.58$ pooled, reaching $+0.82$ on Llama-3-8B. Holding the head and the offset fixed and varying the store size, the margin orders the rate on every model, at 0.86 to 0.97; holding the head and the store size fixed and varying the offset, it orders on every model at 0.72 to 0.89. Within a head, then, the prediction of Theorem 3 holds on all ten models, in both directions of the grid. Exceptions appear only when heads are ranked against one another at a fixed cell, where the coefficient runs 0.29 to 0.71 on eight models and inverts on Qwen2-7B and Qwen2.5-7B, at -0.68 and -0.66. Qwen2-7B and Qwen2.5-7B are the two models whose captured energy is concentrated most heavily in the slow half, at 87% and 88% against at most 78% elsewhere and 64% at the other end, and across ten models the cross-head coefficient tracks that share at Spearman -0.87.

Why the cross-head comparison inverts. At a fixed (L, N) cell the threshold $\sqrt{2\ln(N-1)/d}$ is the same constant for every head, so it cannot change a rank correlation: the cross-head coefficient tests the signal kernel alone, with the interference term absent. That threshold is Theorem 3's floor specialized to a flat profile. Scoring each head against its own floor $\sqrt{2V_{\text{int}}^{(h)}\ln(N-1)}$, with $V_{\text{int}}^{(h)}$ the head's measured interference variance in place of its flat-profile value, leaves the within-head coefficients unchanged to four decimals, as it must, and resolves both inversions: Qwen2-7B and Qwen2.5-7B move from -0.68 and -0.66 to $+0.17$, and the pooled coefficient becomes positive on all ten models, mean 0.54 to 0.70. The correction is not uniformly beneficial: cross-head ρ falls on six models and Yi-6B inverts the other way, 0.33 to -0.33. What it gains tracks slow-half energy share at Spearman $+0.69$, which locates the mechanism: where energy concentrates in the slow half the flat floor understates interference, and where it does not the per-head floor overstates it. We report the flat scoring throughout, as it uses the theorem's own constant.

F.11 Replication at Longer Context

We repeat the capture-side measurements at 32768 tokens on two models: Mistral-7B v0.1, at its declared maximum, where Table 1 gives its spectrum zero coherent bands, and Llama-3.1-8B, at a quarter of its declared context. Protocol, estimator, bin count, nulls, and seeds are unchanged; only the capture length differs. Table 23 sets each measurement against its 8192 value.

Table 23: Capture-side measurements at 32768 against 8192, three-seed means. r_{det} is the median per-head detrended correlation and null the within-layer permutation control, both detrended against the across-head mean curve; the 8192 null is the cross-model band, the 32768 nulls are per-model. Low and high are median long-range mean-logit changes under slow-half and fast-half ablation, viol. the fraction of heads where the fast half costs more, and parity the maximum perplexity difference between cache forms.

Model	length	r_{det}	null	low	high	viol.	parity
Mistral-7B v0.1	8192	0.895	0.13–0.31	6.73	+0.004	0.00%	0.024%
Mistral-7B v0.1	32768	0.924	0.16	4.12	−0.001	0.00%	0.022%
Llama-3.1-8B	8192	0.885	0.13–0.31	8.45	+0.138	0.68%	0.024%
Llama-3.1-8B	32768	0.888	0.21	9.43	+0.085	0.26%	0.017%

The per-head prediction holds at length. Median detrended correlation is 0.924 on Mistral-7B v0.1 and 0.888 on Llama-3.1-8B, against 0.895 and 0.885 at 8192. The controls are unchanged: the within-layer permutation null sits at 0.16 and 0.21, inside the 0.13 to 0.31 band measured at 8192, and the global null is indistinguishable from zero on both models. Each therefore clears its control by 0.76 and 0.67, against 0.62 at 8192. Mistral-7B v0.1 also loses its fixed-lag flags entirely, from 19.3 heads at 8192 to 0: the phase pre-coding those flags detect is a short-range phenomenon.

The long-range channel is cleaner. Ablating the slow half changes the long-range mean logit by 4.12 on Mistral-7B v0.1 and 9.43 on Llama-3.1-8B; ablating the fast half changes it by -0.001 and 0.085. The fast side moves toward zero on both models: it falls by more than a third on Llama-3.1-8B, from 0.138 at 8192, and crosses to negative on Mistral-7B v0.1.

Cache-form equivalence holds at length. Perplexity differs between the identity and RoAM forms by at most 0.022% across both models and all six seeds, against 0.024% at 8192. Seed spread is much wider here than the form difference, 5.36 to 9.74 on Mistral-7B v0.1 and 7.17 to 11.23 on Llama-3.1-8B, because a 32768-token window admits far fewer draws per document. The comparison is paired within seed, so the spread does not enter it.

Spectrum-determined quantities do not move. Alias distances are computed from configuration and must be identical at any capture length. They are: Mistral-7B v0.1's top-two set returns at 138 and Llama-3.1-8B's at 239, matching Table 12 exactly. This is the control that separates predictions fixed by the spectrum from those estimated on captured activations.

Retrieval holds at length. The measurements above are capture-side; we also evaluate retrieval itself at 32768, with the identity cache, the five RULER subtasks, and three seeds. The macro over those subtasks falls from 0.882 at 8192 to 0.853, and the loss is concentrated in variable tracking, from 0.560 to 0.440; single-key, multi-key, multi-value, and ordering move by at most 0.020.

Retrieval and band ablation at 65536. The measurements above establish the band decomposition at 32768 on capture-side quantities, and retrieval at that length on the identity cache alone. We extend both to 65536 on Llama-3.1-8B, whose declared context is 131072, with the same five RULER subtasks, 50 items each, three seeds. Table 24 reports the four arms.

The identity macro continues its decline with length, 0.882 at 8192, 0.853 at 32768, and 0.820 at 65536. The three needle subtasks hold at 0.960 or above; variable tracking and ordering carry nearly all of the decline, and between 32768 and 65536 it is ordering alone, variable tracking sitting at 0.440 at both lengths. Caching in RoAM form moves the macro by $+0.001 \pm 0.004$ and perplexity by less than 0.01%, so the equivalence of Lemma 1 holds at this length as at the others.

The band asymmetry is unchanged at eight times the length at which it was established. Removing the slow half takes every subtask to exactly 0.000 on every seed, fifteen cells in all, and multiplies perplexity by 161. Removing the fast half moves the macro by -0.008 ± 0.012, inside one seed standard deviation of zero, while perplexity moves by 3%.

Table 24: Retrieval at 65536 on Llama-3.1-8B, five RULER subtasks, 50 items each, three-seed means. Ablations remove half the spectrum from the named end, the $f = 1/2$ arms of Figure 3. PPL is full-sequence perplexity; the macro is the unweighted mean of the five subtasks, reported with its spread over seeds. The slow arm reads 0.000 in all fifteen (subtask, seed) cells.

Arm	S-NIAH	MK-NIAH	MV-NIAH	VT	order	macro	PPL
identity	1.000	0.960	0.977	0.440	0.724	0.820 ± 0.009	8.30
RoAM form	1.000	0.967	0.977	0.433	0.728	0.821 ± 0.010	8.30
slow half removed	0.000	0.000	0.000	0.000	0.000	0.000 ± 0.000	1335.5
fast half removed	1.000	0.927	0.958	0.533	0.642	0.812 ± 0.010	8.53

F.12 Replication on a Second Corpus

Table 25 reports the predictions of Corollary 2 on WikiText-103 (Merity et al., 2016), in two forms: with the band cross-covariances re-estimated from WikiText activations, and with the PG-19 coefficients reused unchanged. The first asks whether the functional form holds on a second corpus; the second asks whether the coefficients are properties of the head.

Table 25: WikiText-103 replication, three-seed means. Every quantity is measured on WikiText-103; the PG-19 column is the main-text value being replicated, and Δ is the move between the two corpora, both for the median per-head detrended correlation. Low and high are the median long-range mean-logit changes under slow-half and fast-half ablation; viol. is the fraction of heads where the fast half costs more. η_{sub} and η_{null} are the position-indexed subspace energy and its matched permutation null.

	tuning curves			band ablation			leakage	
Model	WT-103	PG-19	Δ	low	high	viol.	η_{sub}	η_{null}
Mistral-7B v0.1	0.885	0.895	−0.010	6.89	0.004	0.0%	0.047	0.034
Gemma-7B	0.871	0.856	+0.015	4.58	0.004	1.0%	0.016	0.009
SmolLM2	0.924	0.930	−0.006	11.55	−0.013	0.8%	0.118	0.109
Llama-3-8B	0.895	0.910	−0.015	10.11	0.045	0.1%	0.054	0.041
Llama-3.1-8B	0.872	0.885	−0.013	8.85	0.148	0.7%	0.063	0.040
Mistral-7B v0.3	0.872	0.883	−0.011	7.16	0.172	0.0%	0.050	0.033
Qwen2-7B	0.841	0.818	+0.023	7.04	0.177	29.5%	0.046	0.032
Qwen2.5-7B	0.842	0.861	−0.019	6.95	0.184	28.0%	0.046	0.033
Yi-6B	0.934	0.938	−0.004	8.32	0.739	2.5%	0.051	0.037
Granite-3.1	0.921	0.919	+0.002	111.94	3.499	1.7%	0.053	0.040

Tuning curves. The per-head prediction transfers. Median detrended correlation is 0.841 to 0.934 across models, median 0.879, against 0.819 to 0.938 on PG-19 with median 0.894. No model moves by more than 0.023, and the ordering is preserved: Qwen2-7B and Qwen2.5-7B remain the low pair, Yi-6B and SmolLM2 the high. Since the cross-covariances are re-estimated on the new corpus while the spectrum is fixed, the agreement isolates the spectral term as the part that transfers.

Cross-corpus transfer. The predictor of Corollary 2 has no free parameters, but its coefficients, the per-band cross-covariances $\hat{C}_m$, are estimated from activations. To test whether they are properties of the head rather than of the corpus, we score each head's PG-19 prediction against the same head's WikiText-103 measured curve with no re-estimation, and the reverse, under the masked Pearson correlation, head-mean detrending, and five-round within-layer permutation null used above. The head set is the intersection of the two captures, 8992 heads; Granite-3.1's six zero-k layers, live only under the WikiText capture, are dropped. Table 26 reports per-model medians and the shift against the corresponding in-corpus fit. The cross-model median is 0.875 forward and 0.871 backward, against 0.890 and 0.879 in-corpus, and every model clears its null by at least 0.51. Eight models transfer within 0.005 of their in-corpus value forward and 0.02 backward, inside the 0.023 by which the in-corpus fits themselves move between corpora; for these, transfer costs no more than re-estimation. Granite-3.1 loses 0.035 and 0.049. Gemma-7B loses 0.21 and 0.25, to 0.662 and 0.603, while its in-corpus fits sit in the range of the other models (Table 25): the spectrum describes its heads on either corpus, and it is the coefficients that move between them. It is the only model with $d = 256$, so each head carries 128 estimated band phases against 64 elsewhere and 32 on SmolLM2, which has the smallest transfer loss. Whether the larger number of estimated phases is the cause we do not resolve here. On nine of ten models, then, the spectrum fixes the form and the activation distribution fixes coefficients that survive a change of corpus.

Band decomposition. The long-range channel is where PG-19 places it. Ablating the slow half reduces the long-range mean logit by 4.6 to 112 in median magnitude; ablating the fast half changes it by -0.01 to 3.5. We report the two sides separately rather than as a ratio. Direction violations are under 2.5% of heads on eight models. Qwen2-7B and Qwen2.5-7B are the exceptions at 28% and 30%, against 1.8% maximum on PG-19; these are the two models whose band energy is most concentrated in the slow half and the two whose real-key capacity coefficient inverts (Section F.10).

Table 26: Cross-corpus transfer of the per-head predictor: median detrended correlation when PG-19 coefficients are scored on WikiText-103 curves (fwd) and the reverse (bwd), each against its within-layer permutation null. Δ is the shift against the in-corpus fit on the corpus being predicted, so a value near zero means transfer costs no more than re-estimation; n is the number of heads in the intersection.

	PG-19 → WT-103			WT-103 → PG-19			
Model	fwd	null	Δ	bwd	null	Δ	n
Mistral-7B v0.1	0.880	0.206	−0.005	0.882	0.201	−0.013	1024
Gemma-7B	0.662	0.113	−0.207	0.603	0.095	−0.252	448
SmolLM2	0.928	0.135	+0.003	0.917	0.150	−0.013	768
Llama-3-8B	0.898	0.201	+0.002	0.893	0.214	−0.018	1024
Llama-3.1-8B	0.868	0.159	−0.004	0.868	0.178	−0.018	1024
Mistral-7B v0.3	0.869	0.203	−0.003	0.872	0.224	−0.013	1024
Qwen2-7B	0.838	0.172	−0.002	0.810	0.207	−0.008	784
Qwen2.5-7B	0.833	0.167	−0.003	0.850	0.202	−0.014	784
Yi-6B	0.929	0.319	−0.005	0.929	0.325	−0.008	1024
Granite-3.1	0.885	0.191	−0.035	0.870	0.176	−0.049	1088
median	0.875		−0.003	0.871		−0.014	

Aliasing. Alias distances are spectrum-determined: every onset and alias distance reproduces its PG-19 value, seed-invariant on all ten models. Llama-3.1-8B's top-two set returns at 239 on both corpora. This is the control that separates configuration-determined predictions from corpus-estimated ones.

Positional leakage. Section F.8 bounds position dependence from above; a second probe, on blocks drawn from this corpus, gives a lower estimate. For each head we fit a ridge probe from keys to tile index, take the span of its weight matrix as a position-indexed subspace, and measure η_{sub}, the fraction of key energy lying in that span. The chance level is not $\sqrt{r/d}$, since ridge whitening parks unused probe dimensions in low-variance key directions, so we calibrate against a null that permutes tile labels within each content column. Key energy in the subspace is 0.016 to 0.118 against that null at 0.009 to 0.109, separating at 1.7 to 29 standard deviations. The excess is small in absolute terms, and it resolves a position-indexed component rather than bounding one.

F.13 Replication on an Instruction-Tuned Checkpoint

We repeat the capture-side measurements, the band ablation, the two compression methods at the ends of the damage range, and the certificate audit on Llama-3.1-8B-Instruct, the instruction-tuned release of Llama-3.1-8B, on the same ten PG-19 documents with the same estimator, bins, nulls, tasks, and seeds; only the checkpoint differs. Table 27 places each capture-side, capacity, certificate, and leakage quantity beside the base value; Table 28 reports the behavioral arms. All deltas are same-seed differences against the identity cache of the instruction-tuned checkpoint. One task behaves differently under this harness: variable tracking reads 0.007 on the identity cache against 0.560 on the base checkpoint, so the identity retrieval macro, 0.644 against 0.730, is lower by roughly one task; the same-seed deltas do not depend on it.

Table 27: Instruction-tuned replication: capture-side, capacity, certificate, and leakage quantities on Llama-3.1-8B-Instruct beside the base checkpoint, three seeds throughout. Certificate rows give the range over the three reconstruction methods; the base column for those rows is the ten-model range of Table 4. Alias distances for the band sets are spectrum-determined and coincide by construction; the per-head fraction is not, and does not.

	Quantity	Llama-3.1-8B	Llama-3.1-8B-Instruct
Tuning curves	median detrended r	0.886	0.852
	within-layer null	0.159	0.195
	fixed-lag heads flagged	3.1%	2.7%
Band energy	slow-half share, all heads	0.690	0.725
Real keys	Spearman ρ	$+0.81$	$+0.84$
	ρ, vary N	0.95	0.95
	ρ, vary L	0.86	0.86
	ρ, vary head	0.67	0.78
	rate at $N{=}4$	0.72	0.75
	rate at $N{=}1024$	0.14	0.26
Certificate	tightness p_{50}	0.04 to 0.18	0.054 to 0.056
	max ratio	≤ 0.93	≤ 0.49
	rank informativeness ρ	0.18 to 0.71	0.19 to 0.33
Margin screen	certified preserved, three methods	0.000 to 0.025	0.000 to 0.030
	top entry survived, three methods	0.933 to 0.976	0.941 to 0.978
H2O	slots evicted	94%	93.8%
	evicted energy in slow half	69%	69%
Leakage ceiling	median η	0.19	0.20
	peak η	0.26	0.27
	peak layer	13	13
	slow / fast	1.15	1.14
Aliasing	top-2 alias	239	239
	top-4 alias	11693	11693
	heads aliasing at $\epsilon = 0.25$	9.0%	25.4%

Table 28: Behavioral arms on Llama-3.1-8B-Instruct, three-seed means: seven-task retrieval macro, RULER ordering, LongBench passage retrieval, five-task zero-shot macro, PG-19 perplexity, and perplexity relative to the identity cache of the same seed. Zero-shot is not run on the ablation arms, as for the base checkpoints. Base-checkpoint deltas for the two compression rows are -0.08, $+0.00$, $+5.9\%$ (KIVI-2) and -0.65, $+0.00$, $+11.1\%$ (H2O).

Arm	Retrieval	Ordering	Passage retr.	Zero-shot	PPL	PPL ratio
identity	0.644	0.619	0.533	0.685	10.36	1.00
RoAM form	0.637	0.612	0.513	0.685	10.36	1.00
slow, $f = 1/8$	0.010	0.000	0.073		286	28
fast, $f = 1/8$	0.624	0.532	0.480		10.54	1.02
slow, $f = 1/2$	0.000	0.000	0.000		2356	227
fast, $f = 1/2$	0.548	0.439	0.033		12.39	1.20
KIVI-2	0.547	0.549	0.100	0.685	10.93	1.06
H2O	0.054	0.000	0.040	0.685	11.15	1.08

Tuning curves and band energy. The per-head prediction of Corollary 2 holds on the instruction-tuned checkpoint at a median detrended correlation of 0.852 (seed medians 0.847 to 0.861), against 0.886 for the base checkpoint, and clears its within-layer null of 0.195 by 0.66. The loss of 0.034 is of the order of the 0.023 by which the base fits move between corpora and inside the 0.12 spread across the ten base models; the instruction-tuned value sits above Qwen2-7B and Qwen2.5-7B. The slow half holds 0.725 of band energy over all key-value heads and seeds, against 0.690 on the base checkpoint: instruction tuning moves 0.035 of the energy toward the slow half, and the ordering of the two ablation arms is unchanged.

Capacity on real keys. The predicted margin orders the measured rate at Spearman +0.84, against $+0.81$ for the base checkpoint under the same statistic, with the rate declining monotonically in 134 of 144 (head, offset) curves. The decomposition of Section F.10 locates that difference exactly. Within a head the two checkpoints are indistinguishable, at 0.95 against 0.95 varying store size and 0.86 against 0.86 varying offset; the whole of the gain sits in the cross-head comparison, 0.78 against 0.67, above the 0.29 to 0.71 range of every base model in Table 22. Instruction tuning leaves the dependence on store size and distance untouched and makes heads easier to rank against one another.

The long-range channel and deployed compression. Removing an eighth of the slow bands takes six of seven retrieval tasks to 0.000 and multiplies perplexity by 27.6; removing an eighth of the fast bands multiplies perplexity by 1.02 and moves the retrieval macro by 0.02. At $f = 1/2$ the slow-side arm reads 0.000 on every task; the fast-side arm keeps the four needle tasks at or above 0.947 and HotpotQA within 0.01 of identity, lowers ordering to 0.439, and takes passage retrieval to 0.033, as on the base checkpoint, where the same cell reads 0.000 on all three seeds. KIVI-2 and H2O move retrieval by -0.097 ± 0.018 and -0.590 ± 0.023, zero-shot by 0.000 on both, and perplexity by +5.5% and +7.6%: retrieval degrades ahead of zero-shot accuracy after instruction tuning, as before it. Caching in RoAM form moves the retrieval macro by -0.007, inside the seed spread of Table 10.

Certificate, leakage, and aliasing. The certificate holds with zero violations across 12.0 million measured triples, and its tightness, maximum ratio, and rank informativeness lie inside the ten-model ranges of Table 4; the margin screen certifies and observes the same fractions as on base to within 0.01. The leakage ceiling, its peak layer, and its slow-to-fast ratio match the base checkpoint to the precision reported. The alias distances of the band sets are identical to the base row of Table 12, as a shared spectrum requires, but the fraction of heads whose own band support aliases is not: it rises from 9.0% to 25.4% at the registry tolerance, and is higher on the instruction-tuned checkpoint at every tolerance of Table 13. Since the spectrum is shared, the difference lies in which bands each head uses, and its direction matches the shift of band energy toward the slow half reported above: concentrated energy leaves fewer effective bands per head.

G AI ASSISTANCE

Anthropic Claude Opus 4.6 was used for improving the clarity and readability of the writing. The authors take full responsibility for the content.

H REPRODUCIBILITY

Code is provided as supplementary material in the OpenReview submission.