KERNELIZED LINEAR ATTENTION: BREAKING THE CAPACITY WALL WITH SYMMETRIC CONES

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear attention promises constant-time recurrent inference but degrades sharply on associative recall. We formulate attention recall as a spherical-packing problem and introduce Kernelized Linear Attention Activations (KATA), a framework whose feature maps are derived from first principles certifying attention activations on self-dual homogeneous cone. Building on this observation, we show that rank-one positive semi-definite (PSD) features offer a favorable capacity–interference tradeoff. KATA also recovers a parameter-free convex output gate and characterizes associative capacity through the Welch interference floor. On long-range MQAR and repeated-key overwrite, several KATA variants outperform Gated DeltaNet, with parameter counts and state sizes reported alongside accuracy. KATA retains 0.985 MQAR at a $16\times$ out-of-distribution length, approaching the softmax with roughly one quarter of the KV-cache entries. Experiments on 340M-parameter LLMs reveal a feature-dependent fluency trade-off and clarify how RoPE, delta rules, and decay gates interact with feature geometry.

1 INTRODUCTION

Transformer architectures (Vaswani et al., 2017) owe their dominance to softmax attention, whose exponential kernel supports near-perfect associative recall. The price is the $\mathcal{O}(T^2)$ training cost and the $\mathcal{O}(T)$ KV cache at sequence length T, which has motivated a sustained body of linear-time alternatives and fixed-size recurrent states (Katharopoulos et al., 2020; Choromanski et al., 2021; Peng et al., 2021; Sun et al., 2023; Gu & Dao, 2024; Dao & Gu, 2024). Linear attention replaces the infinite feature map of $\exp(\langle \boldsymbol{q}, \boldsymbol{k} \rangle)$ with a finite $\psi : \mathbb{R}^d \to \mathbb{R}^{n_\psi}$ and exposes the recurrent update $\boldsymbol{S}_t = \boldsymbol{S}_{t-1} + \psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top$. The price of constant state is *memory collision*: many key–value bindings must share a single fixed state, and retrieval degrades sharply on associative-recall tasks at long contexts (Schlag et al., 2021; Jelassi et al., 2024; Du et al., 2025).

Recent gated linear RNNs, including DeltaNet (Yang et al., 2024b), Mamba2 (Dao & Gu, 2024), and Gated DeltaNet (Yang et al., 2025b), close part of the gap with learned forget gates, unnormalized linear writes, and content-addressed erase operations. Yet the geometry driving their feature maps remains opaque: what limits recall capacity at a fixed state size, and which post-Transformer designs operate below that limit? Must feature maps be nonnegative? Which finite maps preserve the isotropy of the softmax? Without principled answers, new linear attention architectures sit one ablation away from the existing menagerie of variants.

Core contributions. Memory collision is, at heart, the saturation of a spherical-packing problem. Once we restrict linear attention to match full attention's nonnegative weights and isometric invariance, the Koecher–Vinberg classification (Koecher, 1957; Vinberg, 1963) organizes the admissible feature geometries into three irreducible factors. We study the ordinary real families: the positive orthant, the Lorentz cone, and the symmetric positive semidefinite (PSD) cone. These feature maps form the basis of our **KATA** (Kernelized Linear Attention Activations) framework. The denominator of normalized linear attention emerges natively as a parameter-free, token-conditioned convex gate while the choice of cone determines the interference geometry and, through it, the associative capacity. To compare these geometries, we develop a power signal-to-noise ratio (pSNR) analysis

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.

based on the optimal interference within each cone. We support these predictions with efficient Triton implementations across a wide range of context lengths and with pretraining experiments on 340M-parameter language models at 15B pretraining tokens. Together, the theory and experiments clarify how feature geometry interacts with positional embeddings, gating, the delta rule, and softmax attention.

Working on the unit sphere. Throughout, we analyze query and key directions on the unit sphere. Explicit ℓ_2 normalization enforces this constraint directly. Under RMSNorm, the per-channel gains are absorbed into a learned diagonal bilinear form yielding ℓ_2 normalized directions in a lifted representation (Appendix A.1). This setting covers the query–key normalization used in Qwen3 (Yang et al., 2025a), Gemma 3 (Team et al., 2025), and MiniMax-class (Chen et al., 2026) models, as well as DeltaNet (Yang et al., 2024b) and Gated DeltaNet (Yang et al., 2025b). The capacity bounds therefore describe the normalized address geometry used by these architectures.

2 Reproducing Kernel Hilbert Spaces for Linear Attention

Linear attention (Shen et al., 2021; Katharopoulos et al., 2020) rests on a single algebraic substitution: replacing the exponential kernel of softmax with an inner product of finite features. This substitution allows the quadratic mixing of tokens to collapse into a recurrence on a fixed-size state. We make this substitution explicit and identify the geometric assumptions it imposes on the feature map. Constraining the feature image to a symmetric cone then resolves two design questions typically addressed through ad hoc heuristics: which finite maps are admissible and how normalization produces an output gate.

Throughout, sequences have length T, with $\boldsymbol{q}_t, \boldsymbol{k}_t \in \mathbb{R}^d$ and $\boldsymbol{v}_t \in \mathbb{R}^{d_v}$ denoting the query, key, and value at token t. Their stacked tensors are $\boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{T\times d}$ and $\boldsymbol{V} \in \mathbb{R}^{T\times d_v}$. We write $\psi : \mathbb{R}^d \to \mathbb{R}^{n_\psi}$ for the feature map and $\Psi(\boldsymbol{K})$ for its row-wise application. Causal softmax attention computes, per head:

$$\boldsymbol{z}_t \;=\; \frac{\sum_{i=1}^{t} \exp(\langle \boldsymbol{q}_t, \boldsymbol{k}_i\rangle)\, \boldsymbol{v}_i^\top}{\sum_{i=1}^{t} \exp(\langle \boldsymbol{q}_t, \boldsymbol{k}_i\rangle)}. \tag{1}$$

Substituting the softmax kernel $\exp(\langle \boldsymbol{q}, \boldsymbol{k}\rangle)$ with a finite inner product $\langle \psi(\boldsymbol{q}), \psi(\boldsymbol{k})\rangle$ and using bilinearity yields the linearized form:

$$\boldsymbol{z}_t \;=\; \frac{\psi(\boldsymbol{q}_t)^\top \boldsymbol{S}_t}{\psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_t}, \quad \boldsymbol{S}_t = \sum_{i=1}^{t} \psi(\boldsymbol{k}_i)\boldsymbol{v}_i^\top, \quad \boldsymbol{Z}_t = \sum_{i=1}^{t} \psi(\boldsymbol{k}_i), \tag{2}$$

with state pair $(\boldsymbol{S}_t, \boldsymbol{Z}_t)$ of size $n_\psi(d_v + 1)$ for all t, and additive recurrence $\boldsymbol{S}_t = \boldsymbol{S}_{t-1} + \psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top$, $\boldsymbol{Z}_t = \boldsymbol{Z}_{t-1} + \psi(\boldsymbol{k}_t)$. The normalized readout is defined whenever its denominator $\psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_t$ is positive.

2.1 Symmetric cones

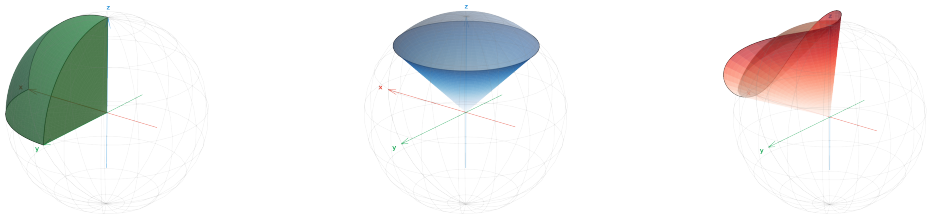
Let $\mathcal{K} \subset \mathbb{R}^{n_\psi}$ be a closed convex cone. Its dual cone is defined as $\mathcal{K}^* = \{\boldsymbol{y} \in \mathbb{R}^{n_\psi} : \langle \boldsymbol{x}, \boldsymbol{y}\rangle \geq 0 \ \forall \boldsymbol{x} \in \mathcal{K}\}$ under the Euclidean inner product. $\mathcal{K}$ is *pointed* if $\mathcal{K} \cap (-\mathcal{K}) = \{0\}$, *full-dimensional* if $\mathrm{int}(\mathcal{K}) \neq \emptyset$, *self-dual* if $\mathcal{K} = \mathcal{K}^*$, and *homogeneous* if its automorphism group acts transitively on $\mathrm{int}(\mathcal{K})$. $\mathcal{K}$ is a *symmetric* cone if it is self-dual and homogeneous. Throughout this paper, we work under the following three assumptions:

Assumption 1 (Self-dual nonnegative geometry)**.** $\psi(\mathbb{R}^d) \subset \mathcal{K} = \mathcal{K}^*$. Consequently, $\langle \psi(\boldsymbol{x}), \psi(\boldsymbol{y})\rangle \geq 0$ for all $\boldsymbol{x}, \boldsymbol{y}$.

Assumption 2 (Isometric invariance)**.** For every $\boldsymbol{U} \in \mathcal{O}(\mathbb{R}^d)$ there exists an inner-product-preserving map $\mathcal{T}_{\boldsymbol{U}}$ on $\mathcal{K}$ such that $\psi(\boldsymbol{U}\boldsymbol{x}) = \mathcal{T}_{\boldsymbol{U}}\psi(\boldsymbol{x})$.

Assumption 3 (Homogeneity)**.** $\mathrm{Aut}(\mathcal{K})$ acts transitively on $\mathrm{int}(\mathcal{K})$.

Where the denominator gate appears. For a fixed query, write $D_t = \psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_t = D_{t-1} + \langle \psi(\boldsymbol{q}_t), \psi(\boldsymbol{k}_t)\rangle$. If $D_t > 0$, the normalized readout is a convex interpolation between the previous normalized readout and the new value, with weights D_{t-1}/D_t and $\langle \psi(\boldsymbol{q}_t), \psi(\boldsymbol{k}_t)\rangle/D_t$. This is the parameter-free gate proved formally in Proposition 6; it comes from normalization, not from an additional state transition.



(a) Positive orthant $\mathbb{R}^3_+$. (b) Lorentz cone $\mathcal{L}^3_+$. (c) PSD cone $\mathbb{S}^2_+$.

Figure 1: **Ordinary real symmetric-cone geometries in $\mathbb{R}^3$.** The PSD cone $\mathbb{S}^2_+$ is linearly isomorphic to the Lorentz cone $\mathcal{L}^3_+$.

2.2 Koecher–Vinberg classification

Isometric invariance from Assumption 2 promotes the softmax identity $\exp(\langle \boldsymbol{U}\boldsymbol{x}, \boldsymbol{U}\boldsymbol{y}\rangle) = \exp(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)$ to the kernel space. On its own, it does not single out $\mathcal{K}$ but guarantees that ψ is $\mathcal{T}_{\boldsymbol{U}}$-equivariant: the input-space isometry $\boldsymbol{U} \in \mathcal{O}(\mathbb{R}^d)$ and the cone-space isometry $\mathcal{T}_{\boldsymbol{U}} \in \mathcal{O}(\mathbb{R}^{n_\psi})|_{\mathcal{K}}$ intertwine through ψ.

Homogeneity in Assumption 3 means that any interior point can be mapped to any other by a cone automorphism. Thus the latent geometry has no privileged interior region or structural bottleneck, a useful symmetry when learning a map into $\mathcal{K}$.

Assumptions 1 and 3 place $\mathcal{K}$ within the scope of the classical Koecher–Vinberg theorem, while Assumption 2 constrains how input rotations act on its factors. The full classification also contains complex Hermitian, quaternionic Hermitian, and exceptional factors. We restrict this work to the ordinary real factors listed below.

Theorem 1 (Koecher–Vinberg decomposition (Koecher, 1957; Vinberg, 1963; Faraut & Korányi, 1994))**.** *Let $\mathcal{K} \subset \mathbb{R}^{n_\psi}$ be a pointed, full-dimensional closed convex cone. If $\mathcal{K}$ is self-dual and homogeneous, then it is a symmetric cone and decomposes uniquely, up to isometry and factor order, as a Cartesian product of irreducible symmetric cones. The irreducible factors studied here are:*

$$\begin{cases} \text{(O)} & \mathbb{R}_+, \\ \text{(L)} & \mathcal{L}^m_+ = \{(\boldsymbol{y}, t) \in \mathbb{R}^{m-1} \times \mathbb{R} : \|\boldsymbol{y}\|_2 \le t\},\ m \ge 3, \\ \text{(S)} & \mathbb{S}^m_+ = \{\boldsymbol{A} \in \mathbb{R}^{m\times m} : \boldsymbol{A} \succeq 0 \text{ and } \boldsymbol{A} = \boldsymbol{A}^\top\},\ m \ge 2. \end{cases}$$

Classical feature maps We can map raw projected keys and queries to the positive orthant via $\mathrm{ReLU}(\boldsymbol{x}) + \epsilon$ for some $\epsilon > 0$, or via $\mathrm{ELU}(\boldsymbol{x}) + 1$ as in linear attention (Katharopoulos et al., 2020). For the Lorentz cone, $\psi^{\mathrm{L}}(\boldsymbol{y}) = (\boldsymbol{y}, \|\boldsymbol{y}\|_2)$ lies on the boundary of $\mathcal{L}^{d+1}_+$ for $\boldsymbol{y} \neq 0$, and Cauchy–Schwarz inequality yields:

$$\langle \psi^{\mathrm{L}}(\boldsymbol{x}), \psi^{\mathrm{L}}(\boldsymbol{y})\rangle = \langle \boldsymbol{x}, \boldsymbol{y}\rangle + \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2 \ge 0.$$

If $\|\boldsymbol{x}\|_2 = 1$, then $\|\psi^{\mathrm{L}}(\boldsymbol{x})\|_2^2 = \langle \psi^{\mathrm{L}}(\boldsymbol{x}), \psi^{\mathrm{L}}(\boldsymbol{x})\rangle = 2\|\boldsymbol{x}\|_2^2 = 2$. Thus $2^{-1/2}\psi^{\mathrm{L}}(\boldsymbol{x})$ lies on the unit sphere.

The PSD cone $\mathbb{S}^m_+$ admits the LDL parameterization detailed in Appendix A.2.

These canonical maps establish nonnegative kernels for each real factor considered here. We next turn from admissible feature geometry to the number of key addresses that each cone can separate.

3 Capacity of Symmetric Cones

The cone classification identifies admissible nonnegative feature geometries, but not how many key addresses a fixed feature dimension can distinguish. Let $\mathcal{K} \subset \mathbb{R}^d$ be a cone and place T unit-norm feature vectors on $\mathcal{K} \cap \mathcal{S}^{d-1}$. We quantify their worst collision by the *interference*:

$$\mu(\mathcal{K}, T) := \min_{\{\boldsymbol{\psi}_1,\dots,\boldsymbol{\psi}_T\} \subset \mathcal{K}\cap\mathcal{S}^{d-1}} \max_{i\neq j} \left|\langle \boldsymbol{\psi}_i, \boldsymbol{\psi}_j\rangle\right|, \tag{3}$$

and define the ϵ-*capacity*:

$$\mathsf{C}_\epsilon(\mathcal{K}) \coloneqq \sup\{T \in \mathbb{N} : \mu(\mathcal{K}, T) \le \epsilon\}. \tag{4}$$

This is precisely the mutual coherence of the feature dictionary (Donoho & Elad, 2003). It isolates the geometry of distinguishable key addresses before specifying values or retrieval dynamics.

Theorem 2 (Welch bound (Welch, 1974))**.** *The d coordinate rays of $\mathbb{R}^d_+$ are pairwise orthogonal, so zero interference is attainable for $T \le d$. For any $T > d$ unit vectors in $\mathbb{R}^d$:*

$$\mu(\mathbb{R}^d, T)^2 \ge \frac{T-d}{d(T-1)}. \tag{5}$$

Consequently, if $\mu \le \epsilon < 1/\sqrt{d}$, then $T \le d(1-\epsilon^2)/(1-d\epsilon^2)$.

Equality conditions and complex-frame extensions are discussed in Appendix A.3.

The positive orthant has small solid angle, but a worst-case packing can live on sparse boundary faces, so volume alone is not a universal capacity obstruction.

The orthant inherits the ordinary Welch floor after its d coordinate rays. For the Lorentz cone $\mathcal{L}^d_+$, a normalized extreme ray has the form $\psi = 2^{-1/2}(\boldsymbol{y}, 1)$ with $\boldsymbol{y} \in \mathcal{S}^{d-2}$, and its kernel is $F_{\mathrm{Lor}}(z) = (1+z)/2$. For $0 \le \epsilon < 1/2$, requiring feature interference at most ϵ forces $\langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle \le -(1-2\epsilon)$. Rankin's bound (see Theorem 7) therefore gives:

$$\mathsf{C}_\epsilon(\mathcal{L}^d_+) \le \min\left\{d,\, 1 + \frac{1}{1-2\epsilon}\right\}.$$

3.1 Rank-one packing into the PSD cone

The LDL construction in Appendix A.2 represents an arbitrary matrix in $\mathbb{S}^p_+$ using:

$$n = \frac{p(p+1)}{2} \tag{6}$$

free coordinates. Directly parameterizing an arbitrary PSD atom would therefore require a learned query/key projection of width n. Packing does not require covering the entire cone: we can project once to $\boldsymbol{u} \in \mathcal{S}^{p-1}$ and apply the deterministic rank-one lift:

$$\psi^{\mathrm{S}}(\boldsymbol{u}) = \boldsymbol{u}\boldsymbol{u}^\top \succeq 0, \qquad \|\boldsymbol{u}\boldsymbol{u}^\top\|_F = 1, \qquad \langle \boldsymbol{u}\boldsymbol{u}^\top, \boldsymbol{v}\boldsymbol{v}^\top \rangle_F = (\boldsymbol{u}^\top \boldsymbol{v})^2.$$

The learned projection emits only p coordinates, while the lifted feature still occupies the n-dimensional symmetric ambient space. Proposition 9 shows that, among unit-Frobenius PSD atoms, rank-one matrices uniquely minimize the expected overlap under a Haar-random relative eigenbasis. This average-case optimality motivates the rank-one family; the following theorem gives its worst-case packing guarantee.

Theorem 3 (PSD rank-one packing)**.** *Let $p \ge 2$, $n = p(p+1)/2$, and $0 \le \epsilon < 1$. Identify $\mathbb{S}^p_+$ with its n-dimensional Euclidean ambient space under the Frobenius inner product. The unit-Frobenius representatives of its extreme rays are rank-one matrices $\boldsymbol{u}\boldsymbol{u}^\top$ with $\boldsymbol{u} \in \mathcal{S}^{p-1}$, and their pairwise inner products satisfy:*

$$\langle \boldsymbol{u}_i \boldsymbol{u}_i^\top, \boldsymbol{u}_j \boldsymbol{u}_j^\top \rangle_F = \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle^2.$$

Hence PSD interference at most ϵ is equivalent to $|\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle| \le \sqrt{\epsilon}$, and a maximal greedy spherical-cap construction yields:

$$\mathsf{C}_\epsilon(\mathbb{S}^p_+) \ge \frac{1}{2}(1-\epsilon)^{-(p-1)/2}. \tag{7}$$

Applying Theorem 2 to the raw vectors shows that PSD interference $\epsilon < 1/p$ forces $T \le p(1-\epsilon)/(1-p\epsilon)$. Rank-one PSD features therefore do not provide exponential capacity below the Welch floor. For any fixed tolerance $\epsilon \in (0,1)$, once $p > 1/\epsilon$ the spherical-cap lower bound in Theorem 3 is exponential in p.

Squaring the raw inner product replaces the mutually obtuse constraint induced by the Lorentz lift with an unsigned near-orthogonal spherical code. Higher even-order extensions are discussed in Appendix A.4.

Deterministic finite dictionaries are given in Appendix D.5.

3.2 SIGNAL-TO-NOISE ANALYSIS

Packing controls the score margin between a matched key and its distractors. Whether that margin is sufficient for retrieval depends on the readout. To isolate key interference, hold the attention coefficients fixed and take independent isotropic values $\boldsymbol{v}_j \sim \mathcal{N}(0, \boldsymbol{I}_{d_v}/d_v)$. Then $\mathbb{E}\|\boldsymbol{v}_j\|^2 = 1$, the cross terms vanish in expectation, and for $\boldsymbol{o}(\boldsymbol{q}) = \sum_j a_j(\boldsymbol{q})\boldsymbol{v}_j$ with matched index $\star$:

$$\mathrm{pSNR}(\boldsymbol{q}) := \frac{\mathbb{E}\|a_\star \boldsymbol{v}_\star\|^2}{\mathbb{E}\|\sum_{j\neq\star} a_j \boldsymbol{v}_j\|^2} = \frac{|a_\star|^2}{\sum_{j\neq\star}|a_j|^2}. \tag{8}$$

We call the readout *ideally retrievable* when $\mathrm{pSNR} > 1$, so the matched value has more power than the aggregate distractors. This criterion gives the expected power ratio under the isotropic-value model. Exact recall also depends on the realized values. Deterministic orthogonal unit values give the same identity when $d_v \geq T$.

Theorem 4 (Idealized Welch-scale retrieval)**.** *Let $p \geq 2$ and $T > p$. Suppose the $T-1$ distractors have squared correlations:*

$$\epsilon^\star = \frac{T-p}{p(T-1)}$$

with the matched query. Then:

$$\mathrm{pSNR}_{\mathrm{Lin}} = \frac{1}{(T-1)\epsilon^\star} = \frac{p}{T-p}, \tag{9}$$

$$\mathrm{pSNR}_{\mathrm{PSD}} = \frac{1}{(T-1)(\epsilon^\star)^2} = \frac{p^2(T-1)}{(T-p)^2}. \tag{10}$$

The criterion $\mathrm{pSNR} > 1$ gives linear retrieval only for $T < 2p$, hence capacity $\Theta(p)$. The PSD map squares distractor interference and extends the idealized threshold to $\Theta(p^2)$, showing how the rank-one lift converts the same address geometry into a larger retrieval regime.

The same pSNR analysis recovers the logarithmic temperature scale for softmax (Chen et al., 2025); the extension and proofs are in Appendices D.8 and D.9.

4 EXPERIMENTS

The theory of Section 3 predicts that associative-recall behavior separates by cone, and the kernel in Appendix H makes the corresponding feature maps cheap to run. We test both predictions on three synthetic tasks that isolate distinct failure modes: multi-query associative recall (MQAR) with controlled length extrapolation, our repeated-key overwrite benchmark, and a long-range induction-head probe. We then scale the recall test to pretrained 340M-parameter language models with a needle-in-a-haystack suite. Only KATA variants instantiated on the PSD cone as described in Appendix H.1 are compared against vanilla linear attention (Katharopoulos et al., 2020), Performer (Choromanski et al., 2021), DeltaNet (Yang et al., 2024b), Gated DeltaNet (Yang et al., 2025b), Mamba2 (Dao & Gu, 2024), and softmax attention. Concrete KATA configurations are named by feature family, KATA-Σg or KATA-Mg. The prefixes *Delta*, *Gated*, and *GatedDelta* add the content erase, multiplicative state gate, or both; for example, GatedDeltaKATA-M2 places PSD features inside the Gated DeltaNet raw-key recurrence. Per-architecture details and reproduction recipes are in Appendices F and F.4. Experimental configurations and the early-stopping criterion are matched across methods.

4.1 MULTI-QUERY ASSOCIATIVE RECALL (MQAR)

We use the canonical Zoology MQAR setup (Arora et al., 2023). Training mixes streams up to K_{train}=64 at $T = 256$; evaluation grids the OOD slice $K \in \{128, 256, 512, 768, 1024\}$ over $T \in \{512, \ldots, 4096\}$, so the largest evaluation slice is $\mathbf{16}\times$ the longest training sequence and the largest training K. We report the best accuracy across the three-point sweep $\mathrm{lr} \in \{10^{-3}, 3\times10^{-3}, 10^{-2}\}$. KATA-Orthant and KATA-Lorentz drop into the same backbone as every other row, varying only the feature map. Table 1 summarizes selected configurations; the full method, head-dimension, and length sweep is in the appendix (Table 5).

Table 1: **MQAR summary.** Selected configurations; accuracy at $8\times$ and $16\times$ the training length. Avg. covers all five OOD lengths in the full appendix table (Table 5). State counts entries, not bytes; *softmax's KV cache is shown at $T = 4096$.

Method	d_h	#params	State	$T = 2048$	$T = 4096$	Avg.
Softmax	64	1,132,928	1,048,576*	1.000	1.000	1.000
GLA	256	5,798,144	65,536	0.902	0.491	0.816
DeltaNet	256	5,520,128	65,536	0.865	0.342	0.748
Gated DeltaNet	256	6,047,750	65,536	0.909	0.419	0.789
KATA-$\Sigma 2$	64	1,134,016	67,584	1.000	0.958	0.991
KATA-M2	64	1,134,016	135,168	1.000	0.970	0.993
KATA-M1	64	1,134,016	266,240	1.000	0.985	0.997
DeltaKATA-M1	64	1,134,272	266,240	1.000	0.999	1.000

In-distribution accuracy saturates for every fixed-state row, so the differences arise entirely on the out-of-distribution slices. The orthant and Lorentz baselines collapse just past $K \approx d_{\text{head}}$, where the linear Welch and Rankin ceilings of Theorems 2 and 7 predict. The PSD variants extend recall into the $16\times$-extrapolation regime. At the largest slice, KATA-M1 retains 0.985 accuracy with a $266{,}240$-entry fixed state, roughly one quarter of softmax's $1{,}048{,}576$ KV-cache entries; DeltaKATA-M1 reaches 0.999 at the same state size. GatedDeltaKATA-M2 exposes the recurrence trade-off: its average overwrite accuracy rises to 0.960, while its MQAR average falls to 0.423. The convex-gate ablation, which caps $\|\boldsymbol{S}_t\|_F$ inside the convex hull of instantaneous bindings, halves the OOD accuracy of the additive form (full table in Appendix F).

4.2 Repeated-key overwrite

We introduce repeated-key overwrite: each of K unique keys is written twice with two *different* values, and the model must retrieve the most recent binding. Training mixes $K \in \{4, 8, 16, 32\}$ at sequence lengths up to 256; evaluation extends to $K \in \{64, 128, 256, 384\}$ with $T \leq 3072$. The backbone is the two-layer Zoology transformer with a width-3 convolutional mixer, parameter-matched across rows.

For a repeated key, an additive state writes both values along the same feature direction and therefore contains a term proportional to $\psi(\boldsymbol{k})(\boldsymbol{v}_{\text{old}} + \boldsymbol{v}_{\text{new}})^\top$. The readout at that address sees their superposition without temporal information.

DeltaKATA-M1 is nearly perfect across the overwrite sweep, while GatedDeltaKATA-M2 reaches 0.960 average overwrite accuracy. Full overwrite results are in Table 6; recurrence details are in Appendices E.2 and H.3.

4.3 Pretraining 340M-parameter language models

To test whether the synthetic results persist under realistic pretraining, we train a family of 340M-parameter models (24 layers, d_{model}=1024, 16 heads) for 15B tokens of SlimPajama (Soboleva et al., 2023) with the TinyLlama tokenizer (Zhang et al., 2024) at a 2048-token context, holding everything fixed and swapping only the token-mixing layer: softmax attention, Gated DeltaNet, KATA-M1 and KATA-M2 (RoPE positions, RMSNorm on $\boldsymbol{q}, \boldsymbol{k}$), DeltaKATA-M1 (KATA-M1 with a DeltaNet content erase), and ℓ_2-norm and short-conv ablations. The models remain broadly comparable on zero-shot accuracy, while perplexity reveals a meaningful fluency spread (Table 4). KATA-M1 trails softmax attention, Gated DeltaNet, and DeltaKATA-M1 on both perplexity datasets; KATA-M2 is higher still. The recall evaluations therefore separate direct association from contextual retrieval: NIAH primarily probes stored bindings, while FDA and SQuAD also require context interpretation and candidate selection. The RoPE analysis is given in Appendix B.3.

Table 2: **In-context recall of pretrained 340M language models.** RULER single-needle accuracy at 1K and 2K (top; 500 samples per length) and Arora'24 cloze accuracy (bottom), both under normalized substring matching. The best and second-best non-softmax primary methods in each column are shown in bold blue and underlined, respectively; diagnostic ablations are unranked.

RULER needle-in-a-haystack						
Method	SNIAH-1 (word)		SNIAH-2 (number)		SNIAH-3 (UUID)	
	1K	2K	1K	2K	1K	2K
Softmax attn.	1.000	0.992	1.000	1.000	0.956	0.930
Gated DeltaNet	0.996	0.954	0.384	0.748	0.004	0.004
KATA-M1	0.994	**0.990**	0.986	0.910	0.758	0.458
DeltaKATA-M1	0.998	0.950	**1.000**	**0.990**	0.908	0.584
KATA-M2	**1.000**	0.974	**1.000**	0.934	**0.910**	**0.616**
Ablations with KATA-M2						
ℓ_2-norm	1.000	0.836	0.982	0.792	0.658	0.224
short-conv, no RoPE	0.956	0.960	0.948	0.926	0.760	0.530

Arora'24 (Based) cloze					
Method	SWDE	FDA	SQuAD	TriviaQA	NQ
Softmax attn.	0.450	0.170	0.160	0.290	0.227
Gated DeltaNet	0.225	0.060	0.293	0.254	0.156
KATA-M1	**0.375**	0.061	0.041	0.258	**0.182**
DeltaKATA-M1	0.370	**0.148**	**0.353**	**0.268**	0.179
KATA-M2	0.338	0.118	0.240	0.256	0.160
Ablations with KATA-M2					
ℓ_2-norm	0.308	0.086	0.259	0.228	0.142
short-conv, no RoPE	0.125	0.074	0.272	0.207	0.088

4.4 NEEDLE-IN-A-HAYSTACK RECALL

Using the same pretrained models, we evaluate the three RULER single-needle tasks S-NIAH-1/2/3 (Hsieh et al., 2024), which hide a word, a 7-digit number, and a 128-bit UUID at a random depth in a distractor haystack and query it with a cloze prompt. The tasks form an increasing-entropy ladder: a UUID is a high-entropy exact string, the regime that most stresses a fixed-size recurrent state. We score the full 500 samples per length with greedy decoding and normalized substring matching at 1K and 2K, both within the 2048-token training horizon. We restrict the main capacity comparison to these in-window lengths because the RoPE-based softmax and KATA models collapse once evaluation moves beyond the training horizon. This protocol isolates retrieval capacity from positional length extrapolation; OOD lengths are reported in Appendix F.

Table 2 places softmax attention at the recall ceiling (≥ 0.93 on every task), while Gated DeltaNet degrades sharply as needle entropy grows: word, number, and UUID recall fall $0.996 \rightarrow 0.384 \rightarrow 0.004$ at 1K. The PSD maps retain much more high-entropy signal. KATA-M1 remains strong on word and number recall, but its UUID accuracy falls from 0.758 at 1K to 0.458 at 2K; KATA-M2 reaches 0.910 and 0.616, respectively.

At equal head width, the geometric packing argument favors KATA-M1's full 64-dimensional factor for pure associative recall. Its recurrent state has 266,240 entries, compared with 135,168 for KATA-M2, so KATA-M2 obtains the stronger hard-needle result with approximately half the state. We interpret this reversal as a learnability effect: KATA-M1 learns one 64-dimensional spherical address factor, while KATA-M2 learns a product of two 32-dimensional factors. Realized recall therefore depends on both the available feature geometry and how readily the projections learn to use it.

The delta erase provides a complementary gain. DeltaKATA-M1 raises UUID recall from 0.758 to 0.908 at 1K and from 0.458 to 0.584 at 2K without changing the KATA-M1 feature width, approaching KATA-M2's 0.910 and 0.616. Two ablations on KATA-M2 isolate the remaining design choices: RMSNorm on $\boldsymbol{q}, \boldsymbol{k}$ improves 2K UUID recall from 0.224 to 0.616, and replacing RoPE with a GDN-style short convolution preserves much of the in-window recall.

Length extrapolation exhibits a different ordering, discussed in Appendix B.1.

4.5 In-context recall on the Based cloze suite

Complementing the controlled NIAH probe (Table 2), we also run the pretrained 340M models on the Based in-context recall suite (Arora et al., 2024), comprising SWDE, FDA, SQuAD, TriviaQA, and NQ. Each task provides a document in context and a cloze query, scored by normalized substring matching on the full dataset. Table 2 reports the results.

SWDE is closest to a direct key–value association. KATA-M1 reaches 0.375, the best fixed-state result, consistent with the pure-recall advantage of its larger feature space. FDA and SQuAD require contextual interpretation and selection among correlated candidates. In this regime, KATA-M1 falls to 0.061 and 0.041, while KATA-M2 reaches 0.118 and 0.240 and DeltaKATA-M1 reaches 0.148 and 0.353. The reversal supports the same learnability interpretation: the single 64-dimensional spherical address factor is harder to train for contextual routing than the product of two 32-dimensional factors. KATA-M1's weaker pretraining perplexity relative to softmax attention, Gated DeltaNet, and DeltaKATA-M1 compounds this difficulty. KATA-M2 improves contextual selection despite its higher perplexity, indicating that factorization and general fluency contribute separately.

Normalized substring matching also rewards verbose generations, so KATA-M1's terse answers are under-credited relative to models that surface several candidate spans. We therefore treat NIAH as the primary capacity result and the Based suite as a joint probe of retrieval, contextual routing, and generation style.

Generation-length statistics and examples are in Appendix B.2.

5 Related Work

We organize prior work along the four axes that Sections 2 and 3 and Appendix H unify: feature geometry, state recurrence, capacity diagnostics, and hardware mapping. For each, we indicate where existing linear-attention designs fall short.

Capacity diagnostics. Zoology (Arora et al., 2023) and MQAR/needle-in-a-haystack benchmarks expose the recall–memory trade-off (Jelassi et al., 2024; Du et al., 2025). Based (Arora et al., 2024) approximates $\exp(\langle \boldsymbol{q}, \boldsymbol{k} \rangle)$ by a degree-2 Taylor map paired with sliding-window attention; LoLA (McDermott et al., 2026) and the GatedDeltaNet-H1/H2 stacks (Yang et al., 2025b) concede the pure linear path and reintroduce a small KV cache. Modern Hopfield networks (Krotov & Hopfield, 2016; Demircigil et al., 2017; Ramsauer et al., 2021; Zhong et al., 2025) establish that exponential energy admits exponentially many stable patterns, and Appendix A.4 places KATA's rank-one PSD map at degree 2 in this hierarchy with $\mathcal{O}(d^2)$ feature state.

An alternative perspective on capacity comes from associative-memory theory, which asks how many patterns remain stable and retrievable under a specified update rule, noise model, and success criterion (Krotov et al., 2025). Higher-order and exponential interactions can raise these dynamical storage thresholds from polynomial to exponential regimes (Krotov & Hopfield, 2016; Demircigil et al., 2017; Lucibello & Mézard, 2024). Our ϵ-capacity measures the quality of the address packing at a fixed interference tolerance. Under the idealized orthogonal-value readout above, that packing directly determines pSNR and hence whether associative recall succeeds. A nonlinear readout can amplify any positive target–distractor margin, while an exact collision remains irreducible. The specified update rule, noise model, and optimization then determine how closely empirical recall realizes this geometric capacity.

Feature geometry. Katharopoulos et al. (2020) replace softmax with $\psi(\boldsymbol{x}) = \mathrm{ELU}(\boldsymbol{x}) + 1$ and exploit the recurrence $\boldsymbol{S}_t = \boldsymbol{S}_{t-1} + \psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top$ to recurse in $\mathcal{O}(n_\psi d_v)$. Performers (Choromanski

et al., 2021), RFA (Peng et al., 2021), cosFormer (Qin et al., 2022), and the fast-weight view of Schlag et al. (2021) use random, trigonometric, or other finite features to approximate $\exp(\langle \boldsymbol{q}, \boldsymbol{k} \rangle)$. When nonnegative attention weights are required, such constructions use orthant-valued maps. The orthant is the simplicial member of the real self-dual cones admitted by the Koecher–Vinberg classification (Koecher, 1957; Vinberg, 1963; Faraut & Korányi, 1994). These works do not analyze the orthant restriction as a capacity bottleneck.

State recurrence. S4 (Gu et al., 2022), RWKV (Peng et al., 2023), RetNet (Sun et al., 2023), and Mamba (Gu & Dao, 2024) generalize the recurrence to a data-dependent transition $\boldsymbol{S}_t = \boldsymbol{A}_t \boldsymbol{S}_{t-1} + \boldsymbol{b}_t \boldsymbol{c}_t^\top$. The State-Space Duality of Mamba2 (Dao & Gu, 2024) restricts $\boldsymbol{A}_t = \alpha_t \boldsymbol{I}$, exposing the algebraic equivalence with causal linear attention but compressing per-token selectivity into a single scalar; the resulting decay is empirically too coarse on retention probes (S-NIAH degrades past 2K tokens (Yang et al., 2025b)). DeltaNet (Yang et al., 2024b) updates the fast-weight matrix $\boldsymbol{S}_t$ by online gradient descent on a squared-error regression objective; Gated DeltaNet (Yang et al., 2025b) adds a forget gate:

$$\boldsymbol{S}_t = \alpha_t (\boldsymbol{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\top) \boldsymbol{S}_{t-1} + \beta_t \boldsymbol{k}_t \boldsymbol{v}_t^\top, \quad \alpha_t, \beta_t \in \mathbb{R}. \tag{11}$$

Proposition 6 shows that normalized linear attention produces the parameter-free convex coefficient $\alpha_t = D_{t-1}/D_t$ at readout while leaving the stored state unchanged. Recurrent gates act directly on memory and therefore provide a distinct control point. When the raw-key substrate in Eq. (11) is constrained to the nonnegative geometry studied here, its Lorentz lift meets the $1/2$-interference Rankin wall of Theorem 7 (Rankin, 1955), giving an $\mathcal{O}(d)$ address regime. Content erase and learned decay can improve overwrite and retention without enlarging the raw feature space. These models tune the recurrence while leaving ψ implicit. Test-time-regression variants (Behrouz et al., 2024; Zuo et al., 2025) operate on the same raw-key substrate; their recurrence mechanisms are complementary to the feature geometry studied here. Separately, Qiu et al. (2025) apply a sigmoid gate after softmax attention, modifying its output rather than a recurrent memory update.

Hardware mapping. Chunkwise parallel training is the standard scaffold for linear-time recurrent attention, with the inter-chunk state propagated by a sequential prefix sum in, for example, the FLA kernels (Yang & Zhang, 2024). For Gated DeltaNet, the non-commutative Householder transition in Eq. (11) requires a $C \times C$ triangular solve per chunk under the WY representation (Schreiber & Van Loan, 1989) and materializes per-token state intermediates in HBM. Mamba2's data-dependent transition instead requires a multiplicative selective scan rather than a vector-space cumulative sum. An additive feature-state recurrence admits a parallel scan that lowers the inter-chunk depth from $\Theta(N_C)$ to $\mathcal{O}(\log N_C)$ (Appendix H.4). KATA combines Koecher–Vinberg geometry with Welch–Rankin packing to certify exponential capacity at a fixed tolerable interference using a finite, hardware-aligned feature map.

6 Discussion and Limitations

KATA characterizes the geometric capacity of attention feature maps, not every mechanism needed for effective attention. Within the real symmetric-cone setting, the framework relates feature geometry to interference and address packing; pSNR connects this geometry to retrieval under its stated assumptions. These results distinguish packing bounds from an idealized retrieval criterion, rather than predicting trained-model quality universally. MQAR tests retention of many associations, whereas our repeated-key overwrite benchmark tests selective replacement of outdated bindings. Increasing address capacity alone cannot separate old and new values written to the same address; DeltaKATA's overwrite results demonstrate the complementary role of the update rule.

Realized performance also depends on learning and contextual fluency. KATA-M2 achieves stronger hard-needle recall than KATA-M1 with approximately half the recurrent state. KATA-M1 is strong on NIAH and direct-copy SWDE but trails KATA-M2 and DeltaKATA-M1 on FDA and SQuAD, which also require context interpretation and candidate selection. KATA-M2 improves contextual selection despite higher perplexity, while generation-length differences complicate substring-based evaluation. These results distinguish storage capacity, memory updates, and broader language-modeling behavior, but do not isolate their causal contributions. KATA provides a framework for the first of these axes, not a complete theory of attention; further directions are discussed in Appendix G.

AI USE STATEMENT

We used generative AI tools to assist with proofreading, checking mathematical proofs, developing GPU kernels, and formatting the manuscript to follow the ICLR 2027 submission guidelines. The authors reviewed all AI-assisted work and take responsibility for the final content, including the text, mathematical claims, proofs, and GPU kernels.

REPRODUCIBILITY STATEMENT

We provide an anonymized code release containing all GPU kernels and the code used to produce the reported results. The theoretical assumptions are stated in Sections 2 and 3, with proofs in Appendices C and D. Experimental protocols and model architectures are detailed in Appendix F, and GPU implementations, hardware benchmarks, and training compute are documented in Appendix H.

REFERENCES

Dimitris Achlioptas. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003. doi: 10.1016/S0022-0000(03)00025-4.

Noga Alon. Perturbed identity matrices have high rank: Proof and applications. *Combinatorics, Probability and Computing*, 18(1–2):3–15, 2009. doi: 10.1017/S0963548307008917.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models, 2023. URL https://arxiv.org/abs/2312.04927.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff, 2024. URL https://arxiv.org/abs/2402.18668.

Somshubhro Bandyopadhyay, P. Oscar Boykin, Vwani Roychowdhury, and Farrokh Vatan. A new proof for the existence of mutually unbiased bases. *Algorithmica*, 34(4):512–528, 2002. doi: 10.1007/s00453-002-0980-7.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time, 2024. URL https://arxiv.org/abs/2501.00663.

Aili Chen, Aonian Li, Baichuan Zhou, Bangwei Gong, Binyang Jiang, Boji Dan, Changhao Zhang, Changqing Yu, Chao Wang, Cheng Ma, Cheng Zhong, Cheng Zhu, Chengjun Xiao, Chengyi Yang, Chengyu Du, Chenyang Zhang, Chi Zhang, Chuangyi Huang, Chunhao Zhang, Chunhui Du, Chunyu Zhao, Congchao Guo, Da Chen, Deming Ding, Dianjun Sun, Dong Li, Dongyu Zhang, Enhui Yang, Fei Yu, Guang Zheng, Guodong Zheng, Guohong Li, Haichao Zhu, Haigang Zhou, Haimo Zhang, Han Ding, Hao Zhang, Haohai Sun, Haolin Lyu, Haonan Lu, Haoyu Wang, Huajie Shi, Huiyang Li, Jiacheng Chen, Jian Zhang, Jiaqi Zhuang, Jiaren Cai, Jiaxin Pan, Jiayao Li, Jiayuan Song, Jichuan Zhang, Jie Wang, Jihao Gu, Jin Zhu, Jingwei Dong, Jingyang Li, Jingyu Zhang, Jingze Zhuang, Jinhao Tian, Jinli Liu, Jinyi Hu, Jun Tao, Jun Zhang, Junbin Ruan, Junhao Xu, Junjie Yan, Junteng Liu, Junxian He, Kang Xu, Ke Ji, Ke Yang, Kecheng Xiao, Keyu Duan, Keyu Li, Le Han, Letian Ruan, Li Yuan, Lianfei Yu, Liheng Feng, Lijie Mo, Lin Li, Linge Du, Lingye Bao, Lingyu Yang, Lingyuan Zhou, Loki, Lu Chen, Lunbin Zeng, Ming Li, Ming Zhong, Mingliang Tao, Mingyuan Chi, Mujie Lin, Nan Hu, Ningxin Chen, Peiyin Zhu, Peng Gao, Pengcheng Gao, Pengfei Li, Penglin Li, Pengyu Zhao, Qibin Ren, Qibing Ren, Qidi Xu, Qihan Ren, Qile Li, Qin Wang, Quanliang Chen, Qunhong Zeng, Rong Tian, Rongxin Guo, Rui Dong, Ruitao Leng, Ruize Zhang, Shanqi Liu, Shaoxiang Chen, Shaoyu Chen, Sheng Jia, Shun Yao, Shuoran Zhao, Shuqi Yu, Sichen Li, Sicheng Pan, Songquan Zhu, Tengfei Li, Tian Xie, Tiancheng Qin, Tianle Li, Tianrun Liang, Wei Liu, Weiqi Xu, Weitao Li, Weixiang Chen, Weiyu Cheng, Weiyu Zhang, Wenhu Chen, Wenqian Zhao, Xiancai Chen, Xiangjun Song, Xiangyuan Wang, Xianzhen Luo, Xiao Luo, Xiao Su, Xiaobo Li, Xiaodong Han, Xiaojie Wu, Xihao Song, Xingyi Han, Xinyu Guan, Xuan Lu, Xun Zou, Xunhao Lai, Xutong Li, Xuyang Shen, Yan Gong, Yan Ma, Yang Jiao, Yang Wang, Yang Xu, Yangsen Wang, Ye Tang, Yicheng Chen, Yihang

Wang, Yinran Qiu, Yiqi Shi, Yiting Guo, Yiwen Huang, Yixuan Wang, Yongyi Hu, Yu Gao, Yu Zhang, Yuan Li, Yuanxiang Ying, Yuanzhen Zhang, Yubo Wang, Yuchen Song, Yufeng Yang, Yuhang Meng, Yuhang Miao, Yuhao Li, Yujie Liu, Yulin Hu, Yunan Huang, Yunji Li, Yunyi Huang, Yusen Zhang, Yusu Hong, Yutao Xie, Yutong Zhang, Yuwen Liao, Yuxuan Shi, Yuze Wenren, Zebin Li, Zehan Li, Zejian Luo, Zeyu Jin, Zeyuan Sun, Zhanpeng Zhou, Zhaochen Su, Zhendong Li, Zhengmao Zhu, Zhengyuan Peng, Zhenhua Fan, Zhi Zhang, Zhichao Xu, Zhiheng Lv, Zhikang Xu, Zhitao He, Zhiwei He, Zhongyuan Li, Zibo Gao, Zijia Wu, Zijian Song, Zijian Zhou, Zijun Sun, Zishan Huang, Ziying Chen, and Ziyue Ge. The minimax-m2 series: Mini activations unleashing max real-world intelligence, 2026. URL https://arxiv.org/abs/2605.26494.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers, 2025. URL https://arxiv.org/abs/2510.05554.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Tri Dao and Albert Gu. Transformers are ssms: generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, 2017. doi: 10.1007/s10955-017-1806-y.

Ronald A. DeVore. Deterministic constructions of compressed sensing matrices. *Journal of Complexity*, 23(4–6):918–925, 2007. doi: 10.1016/j.jco.2007.04.002.

David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, 100(5): 2197–2202, March 2003. doi: 10.1073/pnas.0437847100.

Jusen Du, Weigao Sun, Disen Lan, Jiaxi Hu, and Yu Cheng. Mom: Linear sequence modeling with mixture-of-memories. *arXiv preprint arXiv:2502.13685*, 2025.

Jacques Faraut and Adam Korányi. *Analysis on Symmetric Cones*. Oxford Mathematical Monographs. Oxford University Press, Oxford, 1994. ISBN 9780198534778.

Matthew Fickus and Dustin G. Mixon. Tables of the existence of equiangular tight frames, 2016. URL https://arxiv.org/abs/1504.00253.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

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? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc V. Le. Transformer quality in linear time. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*, 2024.

William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984. doi: 10.1090/conm/026/737400.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Max Koecher. Positivitätsbereiche im $\mathbb{R}^n$. *American Journal of Mathematics*, 79(3):575–596, 1957. doi: 10.2307/2372563.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems 29 (NIPS)*, 2016.

Dmitry Krotov, Benjamin Hoover, Parikshit Ram, and Bao Pham. Modern methods in associative memory, 2025. URL https://arxiv.org/abs/2507.06211.

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

Luke McDermott, Robert W. Heath Jr., and Rahul Parhi. Lola: Low-rank linear attention with sparse caching, 2026. URL https://arxiv.org/abs/2505.23666.

Ivan V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-emnlp.936/.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A. Smith, and Lingpeng Kong. Random feature attention. In *9th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021. URL https://openreview.net/forum?id=QtTKTdVrFBB.

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. In *The Tenth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2022. URL https://openreview.net/forum?id=Bl8CQrx2Up4.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

R. A. Rankin. The closest packing of spherical caps in n dimensions. *Proceedings of the Glasgow Mathematical Association*, 2(3):139–144, 1955.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366. PMLR, 2021. URL http://proceedings.mlr.press/v139/schlag21a.html.

Robert Schreiber and Charles Van Loan. A storage-efficient WY representation for products of Householder transformations. *SIAM Journal on Scientific and Statistical Computing*, 10(1):53–57, 1989. doi: 10.1137/0910005.

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3531–3539, 2021. doi: 10.1109/WACV48630.2021.00357.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. Slimpajama: A 627b token cleaned and deduplicated version of redpajama. https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, 2023.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

E. B. Vinberg. The theory of homogeneous convex cones. *Trudy Moskovskogo Matematicheskogo Obshchestva*, 12:303–358, 1963. URL https://www.mathnet.ru/eng/mmo143.

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

William K. Wootters and Brian D. Fields. Optimal state-determination by mutually unbiased measurements. *Annals of Physics*, 191(2):363–381, 1989. doi: 10.1016/0003-4916(89)90322-9.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

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

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 115491–115522. Curran Associates, Inc., 2024b. doi: 10.52202/079017-3668. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d13a3eae72366e61dfdc7eea82eeb685-Paper-Conference.pdf.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=r8H7xhYPwz.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model, 2024.

Shu Zhong, Mingyu Xu, Tenglong Ao, and Guang Shi. Understanding transformer from the perspective of associative memory, 2025. URL https://arxiv.org/abs/2505.19488.

Yifei Zuo, Yutong Yin, Zhichen Zeng, Ang Li, Banghua Zhu, and Zhaoran Wang. Local linear attention: An optimal interpolation of linear and softmax attention for test-time regression. *arXiv preprint arXiv:2510.01450*, 2025.

A READER'S GUIDE TO THE GEOMETRY

This appendix provides a short map from the attention equations to the geometric claims in the main text. We reserve d for the raw query/key dimension, d_v for the value dimension, and n_ψ for a generic feature dimension. For PSD features, p is the matrix side length and $n := p(p+1)/2$ is the symmetric ambient dimension, so the full PSD map has $n_\psi = n$. The sequence length is T, the chunk size is C, and $N_C := \lceil T/C \rceil$ is the number of chunks. Packing statements define their dictionary size locally.

Why recall becomes a packing problem. Suppose a query $\boldsymbol{q}$ is meant to retrieve the binding at key $\boldsymbol{k}_j$ from the additive state $\boldsymbol{S} = \sum_i \psi(\boldsymbol{k}_i)\boldsymbol{v}_i^\top$. The numerator of the readout is:

$$\psi(\boldsymbol{q})^\top \boldsymbol{S} = \langle \psi(\boldsymbol{q}), \psi(\boldsymbol{k}_j)\rangle \boldsymbol{v}_j^\top + \sum_{i \neq j} \langle \psi(\boldsymbol{q}), \psi(\boldsymbol{k}_i)\rangle \boldsymbol{v}_i^\top .$$

After normalizing features so the matched coefficient is one, every wrong value is weighted by a feature inner product. A fixed recurrent state therefore stores many bindings well only when the stored feature vectors are nearly orthogonal. This is the interference $\mu(\mathcal{K}, T)$ in Equation (3): it asks how many unit feature vectors fit in the allowed cone with all pairwise interference below a target tolerance.

Why cones enter. Normalized linear attention needs nonnegative weights in the denominator and numerator. A self-dual cone $\mathcal{K} = \mathcal{K}^*$ gives a coordinate-free certificate: if $\psi(\boldsymbol{x}), \psi(\boldsymbol{y}) \in \mathcal{K}$, then $\langle \psi(\boldsymbol{x}), \psi(\boldsymbol{y})\rangle \geq 0$. This is weaker than strict coordinatewise positivity. Boundary points are allowed; small ϵ terms are numerical safeguards against a zero denominator, not part of the capacity theorem.

What the three cone factors do. The orthant has d coordinate rays that are exactly orthogonal, but beyond those rays it inherits the ordinary Welch interference floor. The Lorentz lift of a linear key uses $\psi_{\mathrm{Lor}}(\boldsymbol{k}) = 2^{-1/2}(\boldsymbol{k}, 1)$ and turns $z = \langle \boldsymbol{q}, \boldsymbol{k}\rangle$ into $(1+z)/2$; this adds the $1/2$ offset, so pushing below that offset forces the raw keys to be mutually obtuse and invokes Rankin. The PSD rank-one map $\psi_{\mathrm{PSD}}(\boldsymbol{u}) = \boldsymbol{u}\boldsymbol{u}^\top$ turns z into z^2, so antipodal and orthogonal-ish raw keys identifies antipodal raw vectors as the same line, while near-orthogonal raw lines become useful packing directions.

How softmax fits. The exponential kernel is an infinite direct sum of tensor degrees. Theorem 13 shows the cone interpretation of those degrees: even monomials are rank-one PSD rays on tensorized features, while signed odd monomials become nonnegative only when paired with enough even mass into Lorentz-gated PSD atoms. KATA keeps the first capacity-changing PSD degree and implements it with a fixed recurrent state.

A.1 KERNEL METHODS

The exponential dot-product kernel admits the tensor expansion:

$$\exp(\langle \boldsymbol{x}, \boldsymbol{y}\rangle) = \sum_{r=0}^{\infty} \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle^r}{r!} = \sum_{r=0}^{\infty} \left\langle \frac{\boldsymbol{x}^{\otimes r}}{\sqrt{r!}}, \frac{\boldsymbol{y}^{\otimes r}}{\sqrt{r!}} \right\rangle . \tag{12}$$

Thus the canonical RKHS feature map is the infinite direct sum $\phi(\boldsymbol{x}) = \bigoplus_{r\geq 0} \boldsymbol{x}^{\otimes r}/\sqrt{r!}$. Linear attention replaces this infinite feature map by a finite $\psi : \mathbb{R}^d \to \mathbb{R}^{n_\psi}$.

Lemma 5 (Norm of the exponential feature map on the sphere)**.** *For $\tau > 0$, let $K_\tau(\boldsymbol{q}, \boldsymbol{k}) = \exp(\langle \boldsymbol{q}, \boldsymbol{k}\rangle/\tau)$ have canonical feature map:*

$$\phi_\tau(\boldsymbol{x}) = \bigoplus_{r\geq 0} \frac{\boldsymbol{x}^{\otimes r}}{\sqrt{r!\tau^r}}.$$

Then:

$$\|\phi_\tau(\boldsymbol{x})\|_{\mathcal{H}}^2 = K_\tau(\boldsymbol{x}, \boldsymbol{x}) = \exp(\|\boldsymbol{x}\|^2/\tau).$$

In particular, on $\mathcal{S}^{d-1}$ the squared feature norm is the constant $e^{1/\tau}$ and the feature norm is $e^{1/(2\tau)}$.

Proof. The claim follows by evaluating the reproducing kernel on the diagonal: $\|\phi_\tau(\boldsymbol{x})\|_{\mathcal{H}}^2 = \langle \phi_\tau(\boldsymbol{x}), \phi_\tau(\boldsymbol{x}) \rangle_{\mathcal{H}} = K_\tau(\boldsymbol{x}, \boldsymbol{x})$. □

Hence unit-normalizing $\boldsymbol{q}, \boldsymbol{k}$ in input space maps the sphere into a sphere of radius $e^{1/(2\tau)}$ in the exact RKHS, matching the cone-restricted fixed-norm setting used throughout the main text up to a positive scalar.

RMSNorm effect RMSNorm normalizes the input direction and then applies a learned per-channel scale. Its output is therefore not unit norm; nevertheless, its query–key inner product can be represented as a scalar multiple of an inner product between unit vectors in one higher dimension. Ignoring the stabilizing ϵ, let $c_d > 0$ absorb the fixed conversion from RMS normalization to ℓ_2 normalization, together with any fixed attention-logit scaling, and define:

$$\widehat{\boldsymbol{q}} = \frac{\hat{\boldsymbol{q}}}{\|\hat{\boldsymbol{q}}\|}, \qquad \widehat{\boldsymbol{k}} = \frac{\hat{\boldsymbol{k}}}{\|\hat{\boldsymbol{k}}\|}.$$

Define the diagonal gain matrix:

$$\boldsymbol{D}_\gamma \coloneqq c_d \, \mathrm{diag}(\boldsymbol{\gamma}_q \odot \boldsymbol{\gamma}_k).$$

Then:

$$\ell(\hat{\boldsymbol{q}}, \hat{\boldsymbol{k}}) = \widehat{\boldsymbol{q}}^\top \boldsymbol{D}_\gamma \widehat{\boldsymbol{k}} \tag{13}$$

$$= \langle \widehat{\boldsymbol{q}}, \bar{\boldsymbol{k}} \rangle, \qquad \bar{\boldsymbol{k}} \coloneqq \boldsymbol{D}_\gamma \widehat{\boldsymbol{k}}. \tag{14}$$

Choose any $M > 0$ satisfying $M \geq \|\bar{\boldsymbol{k}}\|^2$ for every key under consideration. For a finite dictionary one may take $M = \max_j \|\bar{\boldsymbol{k}}_j\|^2$; a context-independent choice for $\boldsymbol{D}_\gamma \neq 0$ is:

$$M = \|\boldsymbol{D}_\gamma\|_{\mathrm{op}}^2.$$

Define:

$$\widetilde{\boldsymbol{q}} = (\widehat{\boldsymbol{q}}, 0), \qquad \widetilde{\boldsymbol{k}} = \frac{1}{\sqrt{M}} \left(\bar{\boldsymbol{k}}, \sqrt{M - \|\bar{\boldsymbol{k}}\|^2} \right).$$

Both $\widetilde{\boldsymbol{q}}$ and $\widetilde{\boldsymbol{k}}$ are unit vectors in $\mathbb{R}^{d+1}$, and:

$$\ell(\hat{\boldsymbol{q}}, \hat{\boldsymbol{k}}) = \sqrt{M} \, \langle \widetilde{\boldsymbol{q}}, \widetilde{\boldsymbol{k}} \rangle. \tag{15}$$

Thus the RMSNorm-scaled logit is exactly a scalar inverse-temperature factor $\sqrt{M}$ multiplying a cosine similarity in the lifted space. The diagonal gain matrix is absorbed into the lifted key geometry.

A.2 Symmetric cones and Jordan algebras

A subset $\mathcal{K} \subset \mathbb{R}^n$ is a *cone* if $s\boldsymbol{v} \in \mathcal{K}$ for all $s > 0$, $\boldsymbol{v} \in \mathcal{K}$. It is *convex* if $\boldsymbol{u} + \boldsymbol{v} \in \mathcal{K}$ whenever $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{K}$. A closed convex cone is *pointed* if $\mathcal{K} \cap (-\mathcal{K}) = \{0\}$ and *full-dimensional* if its interior is non-empty.

The *dual cone* of $\mathcal{K} \subset \mathbb{R}^n$ under the Euclidean inner product is $\mathcal{K}^* = \{\boldsymbol{y} \in \mathbb{R}^n : \langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq 0 \ \forall \boldsymbol{x} \in \mathcal{K}\}$. The cone is *self-dual* if $\mathcal{K} = \mathcal{K}^*$. Its *automorphism group* is $\mathrm{Aut}(\mathcal{K}) = \{\boldsymbol{A} \in \mathrm{GL}(n, \mathbb{R}) : \boldsymbol{A}(\mathcal{K}) = \mathcal{K}\}$, and $\mathcal{K}$ is *homogeneous* if $\mathrm{Aut}(\mathcal{K})$ acts transitively on the interior $\mathrm{int}(\mathcal{K})$.

The *Koecher–Vinberg theorem* (Koecher, 1957; Vinberg, 1963; Faraut & Korányi, 1994) identifies real symmetric cones with cones of squares in formally real (Euclidean) Jordan algebras, and classifies the irreducible cases as: $\mathbb{R}_+$ (one-dimensional Jordan algebra), $\mathcal{L}_+^m$ for $m \geq 3$ (spin factor), $\mathbb{S}_+^n$ for $n \geq 2$ (real symmetric matrices), $\mathcal{H}_+^n(\mathbb{C})$ (complex Hermitian, $n \geq 2$), $\mathcal{H}_+^n(\mathbb{H})$ (quaternionic Hermitian, $n \geq 2$), and $\mathcal{H}_+^3(\mathbb{O})$ (Albert algebra, exceptional). Our analysis restricts to the first three (real-coordinate Jordan algebras); the others remain valid symmetric cones but require complex-, quaternionic-, or octonionic-valued feature maps.

LDL parameterization. For the PSD cone $\mathbb{S}_+^m$, a classical choice is the LDL parameterization, which represents any $\boldsymbol{A} \in \mathbb{S}_+^m$ as $\boldsymbol{A} = \boldsymbol{L}_A \boldsymbol{D}_A \boldsymbol{L}_A^\top$, where $\boldsymbol{D}_A$ is diagonal and $\boldsymbol{L}_A$ is unit lower triangular. Then:

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{Tr}\big(\boldsymbol{L}_A \boldsymbol{D}_A \boldsymbol{L}_A^\top \boldsymbol{L}_B \boldsymbol{D}_B \boldsymbol{L}_B^\top\big) = \left\| \boldsymbol{D}_B^{1/2} \boldsymbol{L}_B^\top \boldsymbol{L}_A \boldsymbol{D}_A^{1/2} \right\|_{\mathrm{F}}^2.$$

A.3 Fundamental geometric bounds

We collect the four bounds that govern packing on the cone-restricted sphere. Throughout, T is the number of unit vectors and D is the ambient dimension.

Welch bound. For any T unit vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T \in \mathbb{S}^{D-1} \subset \mathbb{R}^D$:

$$\max_{i \neq j} \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle^2 \;\geq\; \frac{T - D}{D(T-1)}.$$

Inverting for T at squared interference ϵ gives $T \leq D(1-\epsilon)/(1-D\epsilon)$ for $\epsilon < 1/D$. As $T \to \infty$, the floor approaches $1/D$. The Welch bound is achieved with equality by Equiangular Tight Frames (ETFs).

The Welch bound (Theorem 2) then gives $\mu(\mathbb{R}_+^{n_\psi}, T)^2 \geq (T - n_\psi)/(n_\psi(T-1))$. At $n_\psi = 128$, requiring $\mu \leq 10^{-2}$ limits the dictionary to $T \lesssim 130$ keys.

Equality conditions and Gerzon bound. For $T > D$, equality in the Welch inequality holds precisely when the vectors form an equiangular tight frame (ETF): tightness minimizes the average squared off-diagonal Gram entry, and equiangularity makes every off-diagonal modulus attain that average. The bound is therefore generally unattained. Real ETFs exist only for special pairs (D, T) and satisfy $T \leq D(D+1)/2$ (Fickus & Mixon, 2016). Over $\mathbb{C}^D$, the corresponding ceiling is D^2, although the representation uses $2D$ real scalars and complex ETF existence is also incomplete. We restrict the present analysis to real cones and leave complex extensions to future work.

Rankin bound. If T unit vectors in $\mathbb{R}^D$ have strictly negative pairwise inner products, then $T \leq D + 1$. At the boundary where the inner products are merely nonpositive, the cross-polytope gives the sharp bound $T \leq 2D$. The Lorentz condition in Theorem 7 is strict: for interference below $1/2$, its spatial components lie in $\mathbb{R}^{d-1}$ and are mutually obtuse, so $T \leq d$.

Alon's combinatorial bound. Alon's rank bound (Alon, 2009) states that for any T unit vectors with $\max_{i \neq j} |\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle| \leq \mu$:

$$D = \Omega\left(\min\left\{ T, \frac{\log T}{\mu^2 \log(1/\mu)} \right\} \right)$$

This universal lower bound leaves only a $\log(1/\mu)$ factor between the best known necessary dimension and the sufficient $D = \mathcal{O}(\mu^{-2} \log T)$ scaling of random spherical codes.

Finite-size comparison. Alon's bound hides an unspecified universal constant, so we do not assign it a numerical value. For $T = 10^5$ and $\mu = 0.05$, the Welch bound requires $D \geq 399$, while the simple all-pairs random-code estimate $D \approx 4 \log T / \mu^2$ gives $D \approx 18{,}421$. The deterministic MUB-derived and DeVore dictionaries below require $D = 802$ and $D = 2{,}209$, respectively.

Table 3: **Constructive packing requirements.** Ambient dimension required for $T = 10^5$ under two interference targets.

Method	$\mu \leq 0.05$	$\mu \leq 0.20$
Welch	$d \geq 399$	$d \geq 25$
Random	$d \approx 18{,}421$	$d \approx 1{,}151$
MUB	$d = 802$ $(s = 401)$	$d = 634$ $(s = 317)$
DeVore	$d = 2{,}209$ $(s = 47)$	$d = 361$ $(s = 19)$

A.4 Dense associative memory hierarchy

The map $\psi^{\mathrm{S}}(\boldsymbol{u}) = \mathrm{vec}(\boldsymbol{u}\boldsymbol{u}^\top)$ squares the inner product: $\langle \psi^{\mathrm{S}}(\boldsymbol{u}_i), \psi^{\mathrm{S}}(\boldsymbol{u}_j) \rangle = \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle^2$. Below the Welch threshold $\epsilon^\star = 1/p$, Theorem 2 gives a finite rational ceiling. For any fixed $\epsilon \in (0, 1)$, once $p > 1/\epsilon$, the greedy spherical-cap construction packs $T \geq \frac{1}{2}(1-\epsilon)^{-(p-1)/2}$ near-orthogonal keys in $p(p+1)/2$ feature coordinates, which grows exponentially in p (Theorem 3).

The cone feature degrees parallel the Krotov–Hopfield Dense Associative Memory (DAM) hierarchy. DAM capacity is defined through an energy, update rule, and stability criterion, whereas our ϵ-capacity counts address directions at fixed interference. The comparison below concerns the tensor-power degree of $F(z) = \langle \psi(\boldsymbol{q}), \psi(\boldsymbol{k}) \rangle$ for $z = \langle \boldsymbol{q}, \boldsymbol{k} \rangle$ and does not identify the two capacity notions.

- **Unnormalized linear attention:** $F_{\mathrm{Lin}}(z) = z$, degree-1 (classical Hopfield), using the raw signed key kernel. Cone nonnegativity is absent.
- **Lorentz lift of the linear kernel:** $F_{\mathrm{Lor}}(z) = \frac{1}{2} + \frac{1}{2}z$. This is the smallest nonnegative cone lift of the degree-1 substrate. Below interference $1/2$, its geometric address capacity is $\mathcal{O}(p)$ by the Rankin wall of Theorem 7.
- **Square-diagonal PSD with rank-one rays:** $F_{\mathrm{PSD}}(z) \propto z^2$, degree-2 (Krotov–Hopfield dense memory). For fixed interference and $p > 1/\epsilon$, the spherical-cap construction of Theorem 3 gives exponentially many address directions. Under the isotropic-value pSNR criterion, aggregate retrieval instead has the $\Theta(p^2)$ threshold of Theorem 4.
- **Softmax:** $F_{\mathrm{soft}}(z) = e^z = \sum_{r \geq 0} z^r / r!$, infinite degree. Theorem 13 shows that its even Taylor degrees are rank-one PSD rays on tensor powers, while its signed odd degrees become positive only after pairing with even mass into Lorentz-gated PSD atoms. The kernel requires the infinite RKHS feature $\phi(\boldsymbol{x}) = \bigoplus_r \boldsymbol{x}^{\otimes r} / \sqrt{r!}$.

Higher even orders. The same p-dimensional projection can generate higher even-order features:

$$\psi_{2r}(\boldsymbol{u}) = \boldsymbol{u}^{\otimes 2r}, \qquad \langle \psi_{2r}(\boldsymbol{u}), \psi_{2r}(\boldsymbol{v}) \rangle = (\boldsymbol{u}^\top \boldsymbol{v})^{2r} \geq 0,$$

whose symmetric ambient dimension is $\binom{p+2r-1}{2r}$. We use the first nontrivial order, $r = 1$: as Appendix H shows, the quadratic expanded state already dominates the hardware budget.

KATA's degree-2 map uses $\mathcal{O}(p^2)$ feature coordinates without adding $\mathcal{O}(p^2 d_v)$ model parameters. The quadratic kernel rematerializes the $p \to p^2$ feature expansion on chip, while the linear-state kernel stores the expanded recurrent state in HBM. Tensor-train factorizations (Oseledets, 2011) are a possible route to higher feature orders without materializing the full tensor; enforcing nonnegativity under such truncations remains open.

B BENCHMARKS RESULTS

Table 4: **Pretraining quality of the 340M models.** For word-level perplexity, lower is better; for zero-shot accuracy, higher is better (`acc_norm` for ARC and HellaSwag, `acc` otherwise; LMBD is LAMBADA). Bold marks the best sub-quadratic result in the columns discussed in the text.

Method	Perplexity ↓		Accuracy ↑						
	LMBD	WikiT	LMBD	ARC-e	ARC-c	HeSw	PIQA	Wino	BoolQ
Softmax attn.	38.6	27.6	0.334	0.412	0.234	0.346	0.645	0.503	0.584
Gated DeltaNet	**40.7**	28.4	0.316	0.393	0.223	**0.349**	0.639	0.519	0.562
KATA-M1	43.5	29.1	**0.319**	0.399	0.222	0.340	**0.642**	0.519	0.585
DeltaKATA-M1	42.5	**28.1**	0.309	**0.403**	0.208	0.343	0.641	0.502	0.597
KATA-M2	51.8	30.0	0.281	0.385	0.230	0.331	0.635	0.512	0.614
Ablations with KATA-M2:									
ℓ_2-norm	57.1	32.2	0.277	0.389	0.230	0.323	0.627	0.499	0.598
short-conv, no RoPE	84.6	34.0	0.233	0.363	0.230	0.313	0.625	0.509	0.615

B.1 NIAH LENGTH EXTRAPOLATION

Length extrapolation exhibits a different ordering. Gated DeltaNet's multiplicative recurrent gate generalizes more reliably beyond 2048 tokens and leads the OOD-length evaluations. Its finite address capacity remains visible on the hard needles within the training horizon: UUID recall is 0.004 at both 1K and 2K despite strong word recall. Strong OOD length generalization and hard-needle recall therefore expose complementary properties of the mixer.

Table 5: **MQAR length extrapolation (full results).** Accuracy (higher is better); Avg. is the unweighted mean over the displayed sequence lengths. Model size and state dimension use the MQAR configuration. ⋆Softmax reports its KV-cache size at T=4096. Within each block, the best and second-best fixed-state results are shown in bold blue and underlined, respectively.

Method	d_h	#params	state (#dim)	MQAR length extrapolation					
				OOD					Avg.
				T=512	T=1024	T=2048	T=3072	T=4096	
Softmax attn.	64	1,132,928	1,048,576⋆	1.000	1.000	1.000	1.000	1.000	1.000
Softmax+RoPE (θ=10^4)	64	1,132,928	1,048,576⋆	0.994	0.658	0.134	0.036	0.012	0.367
Softmax+RoPE (θ=500)	64	1,132,928	1,048,576⋆	0.741	0.046	0.000	0.001	0.001	0.158
RWKV7	64	1,162,880	8,192	0.709	0.224	0.051	0.022	0.013	0.204
Mamba2	64	1,217,548	16,384	0.845	0.413	0.173	0.094	0.054	0.316
GLA	64	1,154,624	8,192	0.900	0.496	0.151	0.067	0.038	0.330
DeltaNet	64	1,134,272	8,192	0.992	0.733	0.249	0.101	0.047	0.424
Gated DeltaNet	64	1,167,878	8,192	**1.000**	**0.944**	**0.515**	**0.279**	**0.170**	**0.582**
RWKV7	128	2,489,600	32,768	0.982	0.670	0.192	0.070	0.032	0.389
Mamba2	128	2,629,132	32,768	0.824	0.336	0.066	0.021	0.008	0.251
GLA	128	2,505,856	32,768	0.993	0.824	0.369	0.178	0.100	0.493
DeltaNet	128	2,432,384	32,768	**1.000**	0.990	0.526	0.244	0.138	0.580
Gated DeltaNet	128	2,565,126	32,768	**1.000**	**0.997**	**0.670**	**0.352**	**0.186**	**0.641**
RWKV7	256	2,489,600	65,536	**1.000**	0.986	0.661	0.332	0.175	0.631
GLA	256	5,798,144	65,536	**1.000**	0.996	0.902	**0.689**	**0.491**	**0.816**
DeltaNet	256	5,520,128	65,536	**1.000**	**1.000**	0.865	0.535	0.342	0.748
Gated DeltaNet	256	6,047,750	65,536	**1.000**	**1.000**	**0.909**	0.618	0.419	0.789
KATA-Σ4	64	1,134,016	17,408	**1.000**	0.997	0.865	0.610	0.407	0.776
KATA-Σ2	64	1,134,016	67,584	**1.000**	**1.000**	**1.000**	0.995	0.958	0.991
KATA-M4	64	1,134,016	69,632	**1.000**	**1.000**	0.982	0.877	0.707	0.913
KATA-M2	64	1,134,016	135,168	**1.000**	**1.000**	**1.000**	0.997	0.970	0.993
KATA-M2+RoPE	64	1,134,016	135,168	0.979	0.384	0.040	0.007	0.003	0.283
DeltaKATA-M2	64	1,134,272	135,168	**1.000**	**1.000**	**1.000**	0.995	0.955	0.990
GatedKATA-M2	64	1,134,276	135,168	**1.000**	**1.000**	**1.000**	0.996	0.965	0.992
GatedDeltaKATA-M2	64	1,134,532	135,168	0.979	0.699	0.258	0.119	0.062	0.423
KATA-M1	64	1,134,016	266,240	**1.000**	**1.000**	**1.000**	0.998	0.985	0.997
DeltaKATA-M1	64	1,134,272	266,240	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**

B.2 Verbosity and fluency examples

Verbosity and fluency. Because normalized substring matching credits any generation that includes the gold span, it rewards length. Table 7 reports the average generated length per model: on SQuAD it ranges from ~ 11 tokens (KATA-M1, which commits to a single terse answer) to ~ 202 (DeltaKATA-M1), an $18.6\times$ spread that mirrors the SQuAD accuracies. Table 8 shows example completions on a shared prompt: every model is fluent, but the terse KATA-M1 model is penalized whenever its short answer is wrong, whereas more verbose models surface the gold span by volume. On the copy-style SWDE task the effect vanishes because all five emit the same short field value. This indicates that the SQuAD/FDA ordering reflects generation style as much as retrieval, a further reason we take NIAH (short, exact needles) as the primary recall measure.

B.3 RoPE and content replacement

Why the delta rule remains useful with RoPE. RoPE assigns repeated occurrences of the same key different phases, so it can in principle resolve overwrite within the 2048-token training window. The resulting separation depends on the attention kernel. Consider an aligned unit query and key in one rotary plane with angular rate ω and relative offset $\Delta = |j - i|$. Relative to $\Delta = 0$, the KATA and softmax factors are:

$$r_{\mathrm{KATA}}(\Delta) = \cos^2(\omega\Delta), \qquad r_{\mathrm{softmax}}(\Delta) = \exp\left(\frac{\cos(\omega\Delta) - 1}{\tau}\right).$$

For $d_{\mathrm{head}} = 64$, RMS-normalized query and key vectors have norm approximately $\sqrt{d_{\mathrm{head}}}$ before learned per-channel gains. Combined with the standard $1/\sqrt{d_{\mathrm{head}}}$ logit scaling, this gives an effective

Table 6: **Repeated-key overwrite: a benchmark introduced in this work.** Accuracy (higher is better); Avg. is the unweighted mean over the displayed key counts. Model size and state dimension use the MQAR configuration. ⋆Softmax reports its KV-cache size at T=4096. Within each block, the best and second-best fixed-state results are shown in bold blue and underlined, respectively.

Method	d_h	#params	state (#dim)	Repeated-key overwrite				
				ID	*OOD*			Avg.
				K=64	K=128	K=256	K=384	
Softmax attn.	64	1,132,928	1,048,576⋆	0.513	0.506	0.499	0.491	0.502
Softmax+RoPE (θ=10^4)	64	1,132,928	1,048,576⋆	0.998	0.890	0.403	0.160	0.613
Softmax+RoPE (θ=500)	64	1,132,928	1,048,576⋆	0.991	0.670	0.140	0.025	0.457
RWKV7	64	1,162,880	8,192	0.820	0.341	0.084	0.030	0.319
Mamba2	64	1,217,548	16,384	0.930	0.556	0.199	0.127	0.453
GLA	64	1,154,624	8,192	0.768	0.317	0.082	0.030	0.299
DeltaNet	64	1,134,272	8,192	0.984	0.691	0.306	0.171	0.538
Gated DeltaNet	64	1,167,878	8,192	**0.998**	**0.853**	**0.384**	**0.217**	**0.613**
RWKV7	128	2,489,600	32,768	0.881	0.376	0.089	0.031	0.344
Mamba2	128	2,629,132	32,768	0.936	0.595	0.224	0.144	0.475
GLA	128	2,505,856	32,768	0.897	0.404	0.105	0.037	0.361
DeltaNet	128	2,432,384	32,768	**1.000**	0.971	0.585	0.347	0.726
Gated DeltaNet	128	2,565,126	32,768	**1.000**	**0.978**	**0.595**	**0.358**	**0.733**
RWKV7	256	2,489,600	65,536	0.938	0.434	0.110	0.038	0.380
GLA	256	5,798,144	65,536	0.927	0.450	0.132	0.050	0.390
DeltaNet	256	5,520,128	65,536	**1.000**	**0.998**	**0.851**	**0.569**	**0.855**
Gated DeltaNet	256	6,047,750	65,536	0.999	0.992	0.726	0.419	0.784
KATA-Σ4	64	1,134,016	17,408	0.517	0.493	0.364	0.226	0.400
KATA-Σ2	64	1,134,016	67,584	0.518	0.505	0.444	0.342	0.452
KATA-M4	64	1,134,016	69,632	0.515	0.484	0.346	0.219	0.391
KATA-M2	64	1,134,016	135,168	0.521	0.509	0.492	0.436	0.489
KATA-M2+RoPE	64	1,134,016	135,168	0.992	0.696	0.087	0.012	0.447
DeltaKATA-M2	64	1,134,272	135,168	**1.000**	0.999	0.992	0.922	0.978
GatedKATA-M2	64	1,134,276	135,168	0.991	0.662	0.193	0.074	0.480
GatedDeltaKATA-M2	64	1,134,532	135,168	**1.000**	0.999	0.983	0.859	0.960
KATA-M1	64	1,134,016	266,240	0.521	0.511	0.503	0.483	0.505
DeltaKATA-M1	64	1,134,272	266,240	**1.000**	**1.000**	**0.999**	**0.998**	**0.999**

Table 7: **Average generation length.** Tokens averaged over 20 examples per task. †256-token cap on NQ.

Method	SWDE	FDA	SQuAD	TriviaQA	NQ†
Softmax attn.	6.5	41.1	54.2	39.5	258
Gated DeltaNet	9.2	27.2	131.8	43.5	258
DeltaKATA-M1	4.5	40.9	**202.2**	46.3	258
KATA-M1	6.5	35.5	**10.9**	72.8	258
KATA-M2	7.0	35.5	182.8	30.4	258

$\tau \approx 1/\sqrt{d_{\text{head}}} = 1/8$ on normalized directions (Appendix A.1). Taking the nominal slow rotary rate $\omega = 1/\theta_{\text{RoPE}}$, with $\theta_{\text{RoPE}} = 10^4$ and $\Delta = 1024$, gives:

$$\omega\Delta = 0.1024, \qquad r_{\text{KATA}} = 0.9896, \qquad r_{\text{softmax}} = 0.9590.$$

The KATA score loses about 1.0% across this offset, while the exponential softmax weight loses about 4.1%. Squaring also removes the sign of the rotated inner product, further weakening phase discrimination. The full RoPE map averages across a spectrum of rates, and this single-plane calculation isolates the effect of the kernel nonlinearity. The delta correction directly updates the content address and can therefore improve recall even inside the training window.

The multiplicative state weights of GatedKATA provide another source of learned positional decay and should strengthen this separation. The 340M-model sweep leaves GatedKATA pretraining and its interaction with RoPE for future work.

Table 8: **Example completions.** One prompt each from SQuAD and SWDE.

SQuAD: cloze "…the NFL team that represented the AFC at Super Bowl 50 was the" **[gold: Denver Broncos]**	
Softmax	New England Patriots. The 2016 NFL season was the 10th season in the NFL…
Gated DeltaNet	New England Patriots. The Patriots won the Super Bowl 50 with a record of 10–1…
DeltaKATA-M1	**Denver Broncos.** The Broncos were led by quarterback Tom Brady, who was named…
KATA-M1	New England Patriots.
KATA-M2	San Francisco 49ers, who defeated the New England Patriots 24–10…
SWDE (structured copy): cloze "…category: Feature year:" **[gold: 1983]**	
all five models	`1983`

C Cone classification and normalization proofs

C.1 The denominator is a parameter-free convex gate

Earlier work on linear attention enforced the nonnegativity in Assumption 1 through the feature map $\psi(\boldsymbol{x}) = \mathrm{ELU}(\boldsymbol{x}) + 1$ (Katharopoulos et al., 2020), whereas Efficient Attention (Shen et al., 2021) omitted the denominator $\psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_t$. Later recurrent architectures instead introduced explicit recurrence-level gates, including the hardware-friendly vector gate of Gated Linear Attention (GLA) (Yang et al., 2024a) and the scalar decay of Mamba2 (Dao & Gu, 2024).

Under Assumption 1, we show that the discarded denominator already supplies a token-conditioned output gate.

Proposition 6 (Output-level gating)**.** *Fix $\boldsymbol{q}_t$ and write:*

$$D_{t-1} = \psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_{t-1}, \qquad c_t = \psi(\boldsymbol{q}_t)^\top \psi(\boldsymbol{k}_t), \qquad D_t = D_{t-1} + c_t.$$

Suppose $D_t > 0$. If $D_{t-1} > 0$, define $\bar{\boldsymbol{z}}_{t|t-1} = \psi(\boldsymbol{q}_t)^\top \boldsymbol{S}_{t-1}/D_{t-1}$. Then:

$$\boldsymbol{z}_t = \alpha_t(\boldsymbol{q}_t)\, \bar{\boldsymbol{z}}_{t|t-1} + \beta_t(\boldsymbol{q}_t)\, \boldsymbol{v}_t^\top, \tag{16}$$

$$\alpha_t = \frac{D_{t-1}}{D_t}, \qquad \beta_t = \frac{c_t}{D_t}, \tag{17}$$

with $\alpha_t, \beta_t \geq 0$ and $\alpha_t + \beta_t = 1$. If $D_{t-1} = 0$, including $t = 1$, then $\boldsymbol{z}_t = \boldsymbol{v}_t^\top$ and the same weights reduce to $(\alpha_t, \beta_t) = (0, 1)$.

The normalized readout is therefore a convex interpolation between the previous readout and the current value, with query-dependent weights and no additional learned parameters. The gate acts at readout, leaving the state update commutative and additive. Gated DeltaNet controls memory through the rank-one state transition (Yang et al., 2025b), which uses extra projections and produces a noncommutative transition.

C.2 Proof of Proposition 6 (output-level gating)

Let $D_{t-1} = \psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_{t-1}$ and $c_t = \psi(\boldsymbol{q}_t)^\top \psi(\boldsymbol{k}_t)$. If $D_{t-1} = 0$, then the nonnegative summands in D_{t-1} are all zero, so $\psi(\boldsymbol{q}_t)^\top \boldsymbol{S}_{t-1} = 0$ as well. Since $D_t = D_{t-1} + c_t > 0$, we have $c_t > 0$ and the normalized readout is exactly $\boldsymbol{v}_t^\top$.

Now suppose $D_{t-1} > 0$. Splitting numerator and denominator of $\boldsymbol{z}_t$:

$$\boldsymbol{z}_t = \frac{\psi(\boldsymbol{q}_t)^\top \boldsymbol{S}_{t-1} + \psi(\boldsymbol{q}_t)^\top \psi(\boldsymbol{k}_t) \boldsymbol{v}_t^\top}{\psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_{t-1} + \psi(\boldsymbol{q}_t)^\top \psi(\boldsymbol{k}_t)}.$$

Multiplying and dividing the first numerator term by $\psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_{t-1}$ yields:

$$\boldsymbol{z}_t = \frac{\psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_{t-1}}{\psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_t}\, \bar{\boldsymbol{z}}_{t|t-1} + \frac{\psi(\boldsymbol{q}_t)^\top \psi(\boldsymbol{k}_t)}{\psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_t}\, \boldsymbol{v}_t^\top.$$

The two numerators sum to $\psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_t$, so $\alpha_t(\boldsymbol{q}_t) + \beta_t(\boldsymbol{q}_t) = 1$. Nonnegativity follows from Assumption 1, because D_{t-1} and c_t are sums of nonnegative feature inner products.

C.3 Canonical-cone classification

Let $\mathcal{K}$ be the closed, pointed, full-dimensional self-dual homogeneous cone that contains $\operatorname{im}\psi$. Self-duality gives $\mathcal{K} = \mathcal{K}^*$, so for any two features $\psi(\boldsymbol{x}), \psi(\boldsymbol{y}) \in \mathcal{K}$:

$$\langle \psi(\boldsymbol{x}), \psi(\boldsymbol{y}) \rangle \geq 0,$$

which certifies Assumption 1. By the Koecher–Vinberg theorem, $\mathcal{K}$ is the cone of squares of a Euclidean Jordan algebra and decomposes uniquely as a Cartesian product of irreducible symmetric cones. Restricting to the ordinary real families considered here leaves $\mathbb{R}_+$, Lorentz cones, and real PSD cones. Products of one-dimensional $\mathbb{R}_+$ factors are orthants. Thus every cone in our scope is represented as a product of orthant, Lorentz, and PSD components. Assumption 2 constrains how rotations of the input act within this product, but does not introduce additional real irreducible cone factors.

D Packing and SNR proofs

D.1 Proof of Theorem 2 (orthant Welch limit)

The d standard basis vectors lie in $\mathbb{R}^d_+$ and are pairwise orthogonal, so zero interference is possible for $N = d$. For $N > d$, apply the Welch bound to any N unit vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N \in \mathbb{R}^d_+$:

$$\max_{i \neq j} \langle \boldsymbol{\psi}_i, \boldsymbol{\psi}_j \rangle^2 \geq \frac{N - d}{d(N-1)}.$$

Because orthant inner products are nonnegative, this is also a lower bound on the squared interference μ^2. If $\mu \leq \epsilon$ and $\epsilon < 1/\sqrt{d}$, rearranging $\epsilon^2 \geq (N-d)/(d(N-1))$ gives $N \leq d(1-\epsilon^2)/(1-d\epsilon^2)$.

D.2 Proof of Theorem 7 (Lorentz interference wall)

Theorem 7 (Lorentz Rankin wall). *Let $0 \leq \epsilon < 1/2$. Any configuration of N unit vectors on the extreme rays of $\mathcal{L}^d_+$ with pairwise interference at most ϵ satisfies:*

$$N \leq \min\left\{ d,\ 1 + \frac{1}{1-2\epsilon} \right\}.$$

Proof. A unit extreme ray of the closure of $\mathcal{L}^d_+$ has the form $\psi = 2^{-1/2}(\boldsymbol{y}, 1)$ with $\boldsymbol{y} \in \mathbb{S}^{d-2}$. Hence:

$$\langle \psi_i, \psi_j \rangle = \frac{1}{2}\langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle + \frac{1}{2}.$$

If $\langle \psi_i, \psi_j \rangle \leq \epsilon < 1/2$, then $\langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle \leq -\alpha$ with $\alpha = 1 - 2\epsilon > 0$. Therefore:

$$0 \leq \|\sum_{i=1}^{N} \boldsymbol{y}_i\|_2^2 = N + 2\sum_{i<j} \langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle \leq N - \alpha N(N-1),$$

which gives $N \leq 1 + 1/\alpha$. Since the pairwise inner products are strictly negative, Rankin's strict spherical-code bound also gives $N \leq d$ for vectors in $\mathbb{R}^{d-1}$ (Rankin, 1955). Combining the two bounds proves the claim. □

Lemma 8 (One-sided spherical-cap bound). *Let $p \geq 2$, let $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{p-1})$, and fix $\boldsymbol{e} \in \mathbb{S}^{p-1}$. For every $a \in [0,1]$:*

$$\Pr[\langle \boldsymbol{u}, \boldsymbol{e} \rangle \geq a] \leq (1-a^2)^{(p-1)/2}. \tag{18}$$

Proof. By rotation invariance, take $\boldsymbol{e}$ to be the north pole and write $\theta = \arccos(a) \in [0, \pi/2]$. With σ denoting normalized surface measure, the cap $C_a(\boldsymbol{e}) = \{\boldsymbol{u} : \langle \boldsymbol{u}, \boldsymbol{e} \rangle \geq a\}$ has measure:

$$\sigma(C_a(\boldsymbol{e})) = \frac{\int_0^\theta \sin^{p-2}\phi \, d\phi}{\int_0^\pi \sin^{p-2}\phi \, d\phi} = \frac{1}{2} I_{1-a^2}\left(\frac{p-1}{2}, \frac{1}{2}\right),$$

where $I_x(\alpha, \beta)$ is the regularized incomplete beta function.

It remains to upper-bound the first ratio. Set $n = p - 2$ and $s = \sin\theta = \sqrt{1 - a^2}$. If $s = 0$, the claim is trivial. Otherwise substitute $\sin\phi = s \sin\psi$ in the numerator. Then:

$$\int_0^\theta \sin^n \phi \, d\phi = s^{n+1} \int_0^{\pi/2} \frac{\sin^n \psi \cos\psi}{\sqrt{1 - s^2 \sin^2 \psi}} \, d\psi.$$

Since $0 \le s \le 1$, we have $\sqrt{1 - s^2 \sin^2 \psi} \ge \sqrt{1 - \sin^2 \psi} = \cos\psi$ for $\psi \in [0, \pi/2]$. Hence:

$$\int_0^\theta \sin^n \phi \, d\phi \le s^{n+1} \int_0^{\pi/2} \sin^n \psi \, d\psi \le s^{n+1} \int_0^{\pi} \sin^n \psi \, d\psi.$$

Dividing by the denominator gives $\sigma(C_a(\boldsymbol{e})) \le s^{n+1} = (1 - a^2)^{(p-1)/2}$. □

D.3 Proof of Theorem 3 (PSD exponential packing)

For $p \ge 2$, the extreme rays of the PSD cone $\mathbb{S}^p_+$ are rank-one matrices $\boldsymbol{u}\boldsymbol{u}^\top$ with $\boldsymbol{u} \in \mathbb{S}^{p-1}$. Their Frobenius inner product is:

$$\langle \boldsymbol{u}_i \boldsymbol{u}_i^\top, \boldsymbol{u}_j \boldsymbol{u}_j^\top \rangle_F = \mathrm{Tr}(\boldsymbol{u}_i \boldsymbol{u}_i^\top \boldsymbol{u}_j \boldsymbol{u}_j^\top) = \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle^2.$$

Thus PSD interference at most ϵ is equivalent to $|\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle| \le \sqrt{\epsilon}$. Let $a = \sqrt{\epsilon}$ and let σ denote normalized surface measure on $\mathbb{S}^{p-1}$.

Now choose lines greedily on $\mathbb{S}^{p-1}$: after selecting $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$, add any $\boldsymbol{u}$ outside the antipodal caps $C_a(\boldsymbol{u}_i) \cup C_a(-\boldsymbol{u}_i)$. When the process is maximal, these antipodal caps cover the sphere; otherwise an uncovered point could be added. By Lemma 8, each selected line covers at most $2(1 - \epsilon)^{(p-1)/2}$ surface measure, so maximality implies:

$$1 \le 2m(1 - \epsilon)^{(p-1)/2} \quad \Rightarrow \quad m \ge \frac{1}{2}(1 - \epsilon)^{-(p-1)/2}.$$

Because each new line is chosen outside all previous antipodal caps, the selected lines satisfy $|\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle| < a$, hence they meet the stated ϵ interference bound after the rank-one PSD map. Adding $\eta \boldsymbol{I}$ to each rank-one matrix and renormalizing moves the construction into the cone interior if a numerical margin is desired, and the pairwise Frobenius inner products converge to the rank-one values as $\eta \downarrow 0$.

D.4 Rank-one PSD atoms minimize Haar-average interference

The ray through $\boldsymbol{u}\boldsymbol{u}^\top$ is extreme. Indeed, if $\boldsymbol{u}\boldsymbol{u}^\top = \boldsymbol{A} + \boldsymbol{B}$ with $\boldsymbol{A}, \boldsymbol{B} \succeq 0$, then for every $\boldsymbol{w} \perp \boldsymbol{u}$, $0 = \boldsymbol{w}^\top \boldsymbol{A} \boldsymbol{w} + \boldsymbol{w}^\top \boldsymbol{B} \boldsymbol{w}$. Positive semidefiniteness gives $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{B}\boldsymbol{w} = 0$, so the ranges of $\boldsymbol{A}$ and $\boldsymbol{B}$ are contained in $\mathrm{span}\{\boldsymbol{u}\}$. Hence $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonnegative multiples of $\boldsymbol{u}\boldsymbol{u}^\top$.

Proposition 9 (Rank-one PSD atoms minimize Haar-average interference)**.** *Let $\boldsymbol{A}, \boldsymbol{B} \succeq 0$ have $\|\boldsymbol{A}\|_F = \|\boldsymbol{B}\|_F = 1$ with eigendecompositions $\boldsymbol{A} = \sum_r a_r \boldsymbol{u}_r \boldsymbol{u}_r^\top$ and $\boldsymbol{B} = \sum_s b_s \boldsymbol{v}_s \boldsymbol{v}_s^\top$, $a_r, b_s \ge 0$. Then $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \sum_{r,s} a_r b_s \langle \boldsymbol{u}_r, \boldsymbol{v}_s \rangle^2 \ge 0$, and under a Haar-random relative eigenbasis $\mathbb{E}\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \mathrm{Tr}(\boldsymbol{A})\,\mathrm{Tr}(\boldsymbol{B})/p \ge 1/p$, with equality only for rank-one matrices.*

Let $\boldsymbol{A} = \sum_r a_r \boldsymbol{u}_r \boldsymbol{u}_r^\top$ and $\boldsymbol{B} = \sum_s b_s \boldsymbol{v}_s \boldsymbol{v}_s^\top$ be spectral decompositions with $a_r, b_s \ge 0$. Then:

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \mathrm{Tr}\left(\sum_{r,s} a_r b_s \boldsymbol{u}_r \boldsymbol{u}_r^\top \boldsymbol{v}_s \boldsymbol{v}_s^\top \right) = \sum_{r,s} a_r b_s \langle \boldsymbol{u}_r, \boldsymbol{v}_s \rangle^2.$$

All coefficients are nonnegative, so high-rank PSD features average many rank-one interactions without cancellation. If the relative eigenbasis is Haar-random, then $\mathbb{E}\langle \boldsymbol{u}_r, \boldsymbol{v}_s \rangle^2 = 1/p$ for every pair (r, s), hence:

$$\mathbb{E}\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \frac{1}{p} \left(\sum_r a_r \right) \left(\sum_s b_s \right) = \frac{\mathrm{Tr}(\boldsymbol{A})\,\mathrm{Tr}(\boldsymbol{B})}{p}.$$

Since $\|\boldsymbol{A}\|_F^2 = \sum_r a_r^2 = 1$, Cauchy–Schwarz gives $\mathrm{Tr}(\boldsymbol{A}) = \sum_r a_r \geq 1$, with equality iff exactly one eigenvalue is nonzero. The same holds for $\boldsymbol{B}$, so the expected interference is at least $1/p$, with equality only when both matrices are rank one.

The Cholesky identity used in the main text follows from cyclicity of trace: if $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$ and $\boldsymbol{B} = \boldsymbol{M}\boldsymbol{M}^\top$, then:

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \mathrm{Tr}(\boldsymbol{L}\boldsymbol{L}^\top \boldsymbol{M}\boldsymbol{M}^\top) = \mathrm{Tr}(\boldsymbol{L}^\top \boldsymbol{M}\boldsymbol{M}^\top \boldsymbol{L}) = \|\boldsymbol{L}^\top \boldsymbol{M}\|_F^2.$$

Thus dense factors introduce all cross-column products. A single flat-vector inner product omits these interactions.

D.5 Constructive packing

We give two deterministic finite dictionaries with explicit interference guarantees. The first realifies complete sets of complex mutually unbiased bases (MUBs), producing a dense dictionary whose size is quadratic in its real ambient dimension.

Theorem 10 (MUB-derived real dictionary)**.** *For any prime power $s = \ell^k$, realification of a complete set of complex MUBs gives $T = s(s+1)$ unit vectors in $\mathbb{R}^d$, where $d = 2s$, with:*

$$\mu(\mathbb{R}^d, T) \leq \frac{1}{\sqrt{s}} = \sqrt{\frac{2}{d}}.$$

This construction is useful when $T = \Theta(d^2)$ is sufficient and a uniform deterministic worst-case guarantee matters more than sparsity. The prime-power condition restricts the available dimensions. The full construction is given in Appendix D.6. When higher interference is tolerated, DeVore's polynomial construction provides a larger and sparser dictionary.

Theorem 11 (DeVore polynomial construction)**.** *For any prime power s and integer $0 \leq r < s$, there is a sparse deterministic construction of $T = s^{r+1}$ unit vectors in $\mathbb{R}^d$, where $d = s^2$, with maximum interference $\mu \leq r/s$.*

This construction is preferable when sparse storage or fast overlap computation matters more than the tighter interference of the MUB-derived dictionary. The proof is in Appendix D.7.

The Johnson–Lindenstrauss (JL) lemma gives a related randomized asymptotic statement: T prescribed points can be embedded into $d = \mathcal{O}(\eta^{-2} \log(T/\delta))$ dimensions while preserving all pairwise distances within $1 \pm \eta$ with failure probability at most δ (Johnson & Lindenstrauss, 1984; Achlioptas, 2003). The same concentration argument yields random spherical dictionaries with interference at most μ in $d = \mathcal{O}(\mu^{-2}(\log T + \log(1/\delta)))$ dimensions, or equivalently an achievable $T = \exp(\Omega(\mu^2 d))$ at fixed success probability. These asymptotic random-projection guarantees are often loose at the dimensions of individual attention heads. Deterministic constructions give exact finite dictionaries and worst-case interference guarantees in the regimes where they apply. The bounds and finite-size comparison are collected in Appendix A.3 and Table 3. These constructions illustrate the available finite regimes, and better real spherical packings may exist. Determining which packings are best suited to attention heads is left to future work.

D.6 MUB-derived real dictionary

Theorem 10 (MUB-derived real dictionary)**.** *For any prime power $s = \ell^k$, realification of a complete set of complex MUBs gives $T = s(s+1)$ unit vectors in $\mathbb{R}^d$, where $d = 2s$, with:*

$$\mu(\mathbb{R}^d, T) \leq \frac{1}{\sqrt{s}} = \sqrt{\frac{2}{d}}.$$

Proof. We first write the construction explicitly when $s = p$ is an odd prime. Let $\omega = \exp(2\pi i/p)$ and let $\boldsymbol{e}_0, \ldots, \boldsymbol{e}_{p-1}$ be the standard basis of $\mathbb{C}^p$. The first basis is $\{\boldsymbol{e}_0, \ldots, \boldsymbol{e}_{p-1}\}$. For each $k \in \mathbb{F}_p$ and $m \in \mathbb{F}_p$, define:

$$\boldsymbol{u}_{k,m} = \frac{1}{\sqrt{p}} \sum_{j=0}^{p-1} \omega^{kj^2 + mj} \boldsymbol{e}_j,$$

with all exponents evaluated modulo p. This gives p additional bases, one for each $k \in \mathbb{F}_p$.

Within a fixed k, the inner product of two vectors is:

$$\langle \boldsymbol{u}_{k,m}, \boldsymbol{u}_{k,m'} \rangle_{\mathbb{C}} = \frac{1}{p} \sum_{j=0}^{p-1} \omega^{(m'-m)j},$$

which is 1 if $m = m'$ and 0 otherwise. Thus each fixed-k collection is an orthonormal basis. A standard basis vector has inner-product modulus $1/\sqrt{p}$ with every $\boldsymbol{u}_{k,m}$, since every coordinate of $\boldsymbol{u}_{k,m}$ has modulus $1/\sqrt{p}$. Finally, for $k \neq k'$:

$$\langle \boldsymbol{u}_{k,m}, \boldsymbol{u}_{k',m'} \rangle_{\mathbb{C}} = \frac{1}{p} \sum_{j=0}^{p-1} \omega^{(k'-k)j^2+(m'-m)j}.$$

The quadratic coefficient is nonzero, so the Gauss-sum identity gives this sum modulus $\sqrt{p}$; hence the inner-product modulus is $1/\sqrt{p}$.

The same finite-field construction, with the appropriate trace character, gives complete sets of $s+1$ mutually unbiased bases in $\mathbb{C}^s$ for every prime power $s = \ell^k$ (Wootters & Fields, 1989; Bandyopadhyay et al., 2002). Write these bases as $\mathcal{B}_a = \{\boldsymbol{u}_{a,1}, \ldots, \boldsymbol{u}_{a,s}\}$ for $a = 1, \ldots, s+1$. They satisfy:

$$|\langle \boldsymbol{u}_{a,i}, \boldsymbol{u}_{b,j} \rangle_{\mathbb{C}}| = \begin{cases} 0, & a = b,\ i \neq j, \\ 1/\sqrt{s}, & a \neq b, \end{cases}$$

with all vectors unit norm. Thus the complex construction already contains $m = s(s+1)$ unit vectors with maximum complex interference $1/\sqrt{s}$.

To obtain real vectors, apply the realification map $R : \mathbb{C}^s \to \mathbb{R}^{2s}$, $R(\boldsymbol{a} + i\boldsymbol{b}) = (\boldsymbol{a}, \boldsymbol{b})$. This map preserves norms: $\|R(\boldsymbol{u})\|_2^2 = \|\Re \boldsymbol{u}\|_2^2 + \|\Im \boldsymbol{u}\|_2^2 = \|\boldsymbol{u}\|_2^2$. Moreover, for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^s$:

$$\langle R(\boldsymbol{u}), R(\boldsymbol{v}) \rangle_{\mathbb{R}} = \Re \langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathbb{C}}.$$

Hence $|\langle R(\boldsymbol{u}), R(\boldsymbol{v}) \rangle_{\mathbb{R}}| \leq |\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathbb{C}}|$. Realifying the $s(s+1)$ MUB vectors therefore gives $m = s(s+1)$ unit vectors in $\mathbb{R}^{2s}$ with maximum absolute interference at most $1/\sqrt{s}$. Since $d = 2s$, this is $1/\sqrt{s} = \sqrt{2/d}$. □

D.7 DeVore construction

Theorem 11 (DeVore polynomial construction)**.** *For any prime power s and integer $0 \leq r < s$, there is a sparse deterministic construction of $T = s^{r+1}$ unit vectors in $\mathbb{R}^d$, where $d = s^2$, with maximum interference $\mu \leq r/s$.*

Proof. For a prime p, take $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ and index the rows by pairs $(x, y) \in \mathbb{F}_p^2$. The same construction works over the finite field $\mathbb{F}_s$ for any prime power s, so we write it in that notation. Following DeVore's finite-field construction (DeVore, 2007), index the coordinates of $\mathbb{R}^{s^2}$ by $(x, y) \in \mathbb{F}_s^2$ and choose one polynomial for every coefficient vector $\boldsymbol{a} = (a_0, \ldots, a_r) \in \mathbb{F}_s^{r+1}$:

$$P_{\boldsymbol{a}}(x) = a_0 + a_1 x + \cdots + a_r x^r.$$

There are s^{r+1} such coefficient vectors. The condition $r < s$ ensures that distinct coefficient vectors define distinct functions on $\mathbb{F}_s$: if $P_{\boldsymbol{a}} = P_{\boldsymbol{b}}$ at all s field elements, then $P_{\boldsymbol{a}} - P_{\boldsymbol{b}}$ is a nonzero polynomial of degree at most r with more than r roots, impossible over a field.

For each polynomial P, define $\boldsymbol{v}_P \in \mathbb{R}^{s^2}$ by:

$$(\boldsymbol{v}_P)_{(x,y)} = \begin{cases} 1/\sqrt{s}, & y = P(x), \\ 0, & y \neq P(x). \end{cases}$$

The graph $\{(x, P(x)) : x \in \mathbb{F}_s\}$ has exactly s coordinates, so $\|\boldsymbol{v}_P\|_2^2 = s \cdot (1/s) = 1$.

For two distinct polynomials P, Q, the supports of $\boldsymbol{v}_P$ and $\boldsymbol{v}_Q$ overlap exactly at the field elements where $P(x) = Q(x)$. Therefore:

$$\langle \boldsymbol{v}_P, \boldsymbol{v}_Q \rangle = \frac{1}{s} \left| \{x \in \mathbb{F}_s : P(x) = Q(x)\} \right|.$$

The polynomial $P-Q$ is nonzero and has degree at most r. By the standard root bound for univariate polynomials over a field, it has at most r roots in $\mathbb{F}_s$. Hence $\langle \boldsymbol{v}_P, \boldsymbol{v}_Q \rangle \le r/s$. This gives $m = s^{r+1}$ unit vectors in $\mathbb{R}^{s^2}$ with maximum interference at most r/s. □

D.8 Proof of Theorem 4 (idealized Welch-scale retrieval)

Proof. Normalize the matched score to one. For linear features, each distractor has score magnitude $\sqrt{\epsilon^\star}$ and therefore contributes power $\epsilon^\star$. For rank-one PSD features, the score is the squared raw correlation, so each distractor contributes power $(\epsilon^\star)^2$. Summing over $T-1$ distractors gives:

$$\mathrm{pSNR}_{\mathrm{Lin}} = \frac{1}{(T-1)\epsilon^\star}, \qquad \mathrm{pSNR}_{\mathrm{PSD}} = \frac{1}{(T-1)(\epsilon^\star)^2}.$$

Substituting $\epsilon^\star = (T-p)/(p(T-1))$ yields the two exact expressions in the theorem.

Finally, $\mathrm{pSNR}_{\mathrm{Lin}} > 1$ is equivalent to $T < 2p$. For PSD features, setting $\mathrm{pSNR}_{\mathrm{PSD}} = 1$ gives:

$$T^2 - (p^2+2p)T + 2p^2 = 0,$$

whose larger root is $p^2 + 2p - 2 + \mathcal{O}(p^{-1})$, giving the stated $\Theta(p^2)$ threshold. □

D.9 Proof of Proposition 12 (softmax sharpening)

We can leverage the comparison in Theorem 4 to expose what changes when the readout is activated through softmax with temperature τ. This yields Proposition 12 and recovers, from a complementary retrieval perspective, the critical logarithmic scale identified for long-context softmax dynamics by Chen et al. (2025). In our framework the law follows directly from Welch-scale interference and the pSNR criterion: the matched score must be amplified logarithmically to remain above the aggregate power of a growing number of distractors.

Proposition 12 (Softmax sharpening at the Welch scale)**.** *Under the setup of Theorem 4, let $\mu^\star = \sqrt{\epsilon^\star}$ and take the harmful distractor sign $+\mu^\star$. Temperature-τ softmax satisfies:*

$$\mathrm{pSNR}_{\mathrm{Soft}} = \frac{\exp(2(1-\mu^\star)/\tau)}{T-1}. \tag{19}$$

Any finite target length with $\mu^\star < 1$ interference is retrievable whenever:

$$\frac{1}{\tau} > \frac{\log(T-1)}{2(1-\mu^\star)}. \tag{20}$$

Proof. Write $z_j = \langle \boldsymbol{q}, \boldsymbol{k}_j \rangle$. Because $\boldsymbol{q}$ and $\boldsymbol{k}_j$ have unit norm:

$$z_j + 1 = 2\langle \psi_{\mathrm{Lor}}(\boldsymbol{q}), \psi_{\mathrm{Lor}}(\boldsymbol{k}_j) \rangle \ge 0, \qquad \psi_{\mathrm{Lor}}(\boldsymbol{x}) = 2^{-1/2}(\boldsymbol{x}, 1).$$

Adding a common bias does not change softmax:

$$\frac{e^{(z_j+b)/\tau}}{\sum_\ell e^{(z_\ell+b)/\tau}} = \frac{e^{z_j/\tau}}{\sum_\ell e^{z_\ell/\tau}}.$$

At the matched key $z_\star = 1$, while the harmful Welch-scale distractors have $z_j = \mu^\star = \sqrt{\epsilon^\star}$. After canceling the common bias factor, their unnormalized scores are:

$$s_\star = e^{1/\tau}, \qquad s_d = e^{\mu^\star/\tau}.$$

The partition is common to every coefficient and cancels from pSNR. Under the independent isotropic-value model of Equation (8):

$$\mathrm{pSNR}_{\mathrm{Soft}} = \frac{s_\star^2}{(T-1)s_d^2} = \frac{e^{2(1-\mu^\star)/\tau}}{T-1}.$$

Therefore $\mathrm{pSNR}_{\mathrm{Soft}} > 1$ is exactly

$$T - 1 < \exp\left(\frac{2(1-\mu^\star)}{\tau}\right). \tag{21}$$

□

When $T \gg p$, $\mu^\star = \sqrt{(T-p)/(p(T-1))} \to 1/\sqrt{p}$, which gives:

$$T_{\mathrm{Soft}}(p,\tau) \approx 1 + \exp\left(\frac{2(1-1/\sqrt{p})}{\tau}\right). \tag{22}$$

It is therefore ideally retrievable precisely when:

$$T - 1 < \exp\left(\frac{2(1-\mu^\star)}{\tau}\right). \tag{23}$$

For $T \gg p$ and Welch-saturated packing, this gives the approximate capacity:

Solving the same inequality for $1/\tau$ gives Equation (20).

D.10 Softmax Taylor cone decomposition

Theorem 13 (Softmax Taylor components factor through cone atoms)**.** *Let $z = \langle \boldsymbol{q}, \boldsymbol{k} \rangle$ for unit vectors $\boldsymbol{q}, \boldsymbol{k} \in \mathbb{S}^{d-1}$, and define $\boldsymbol{t}_r(\boldsymbol{x}) = \mathrm{vec}(\boldsymbol{x}^{\otimes r})$. Every even Taylor monomial factors through rank-one PSD atoms:*

$$z^{2r} = \langle \boldsymbol{t}_r(\boldsymbol{q})\boldsymbol{t}_r(\boldsymbol{q})^\top, \boldsymbol{t}_r(\boldsymbol{k})\boldsymbol{t}_r(\boldsymbol{k})^\top \rangle_F.$$

The odd-degree terms become nonnegative after Lorentz gating, and the exponential kernel admits the decomposition:

$$e^z = \sum_{r=0}^{\infty} \frac{z^{2r}(1+z)}{(2r+1)!} + \sum_{r=1}^{\infty} \frac{2r\, z^{2r}}{(2r+1)!}.$$

Proof. For unit $\boldsymbol{q}, \boldsymbol{k}$ and $\boldsymbol{t}_r(\boldsymbol{x}) = \mathrm{vec}(\boldsymbol{x}^{\otimes r})$:

$$\langle \boldsymbol{t}_r(\boldsymbol{q}), \boldsymbol{t}_r(\boldsymbol{k}) \rangle = \langle \boldsymbol{q}, \boldsymbol{k} \rangle^r = z^r.$$

Thus:

$$z^{2r} = \langle \boldsymbol{t}_r(\boldsymbol{q}), \boldsymbol{t}_r(\boldsymbol{k}) \rangle^2 = \langle \boldsymbol{t}_r(\boldsymbol{q})\boldsymbol{t}_r(\boldsymbol{q})^\top, \boldsymbol{t}_r(\boldsymbol{k})\boldsymbol{t}_r(\boldsymbol{k})^\top \rangle_F,$$

which is the inner product of two rank-one rays in the PSD cone over the r-fold tensor feature space. An odd monomial satisfies $z^{2r+1} < 0$ at $\boldsymbol{k} = -\boldsymbol{q}$ and therefore violates nonnegativity on the full sphere.

Let $\ell(\boldsymbol{x}) = (\boldsymbol{x}, 1)$. For $\boldsymbol{x} \in \mathbb{S}^{d-1}$, $\ell(\boldsymbol{x}) \in \partial \mathcal{L}_+^{d+1}$ and:

$$\langle \ell(\boldsymbol{q}), \ell(\boldsymbol{k}) \rangle = 1 + z.$$

Consequently:

$$z^{2r}(1+z) = \langle \boldsymbol{t}_r(\boldsymbol{q})\boldsymbol{t}_r(\boldsymbol{q})^\top, \boldsymbol{t}_r(\boldsymbol{k})\boldsymbol{t}_r(\boldsymbol{k})^\top \rangle_F \langle \ell(\boldsymbol{q}), \ell(\boldsymbol{k}) \rangle,$$

the product of a rank-one PSD-ray kernel and a Lorentz-ray kernel. Equivalently, define:

$$\Phi_r(\boldsymbol{x}) = \left(\boldsymbol{t}_r(\boldsymbol{x})\boldsymbol{t}_r(\boldsymbol{x})^\top\right) \otimes \ell(\boldsymbol{x}).$$

Then $\langle \Phi_r(\boldsymbol{q}), \Phi_r(\boldsymbol{k}) \rangle = z^{2r}(1+z)$ under the tensor-product inner product, making the "multiplication" a standard feature-map tensoring operation.

It remains to check that the Taylor coefficients can be regrouped this way. Absolute convergence on $[-1, 1]$ permits rearrangement, and:

$$\sum_{r=0}^{\infty} \frac{z^{2r}(1+z)}{(2r+1)!} + \sum_{r=1}^{\infty} \frac{2r\, z^{2r}}{(2r+1)!}$$

has odd coefficient $1/(2r+1)!$ on z^{2r+1} and even coefficient:

$$\frac{1}{(2r+1)!} + \frac{2r}{(2r+1)!} = \frac{1}{(2r)!}$$

on z^{2r} for $r \geq 1$, while the constant term is 1. This is exactly the Taylor series of e^z.

□

E KATA VARIANTS AND RECURRENCES

E.1 VARIANT TAXONOMY

We use three explicit KATA variants:

Plain KATA (additive recurrence, no gate). $\boldsymbol{S}_t = \boldsymbol{S}_{t-1} + \psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top$, $\boldsymbol{Z}_t = \boldsymbol{Z}_{t-1} + \psi(\boldsymbol{k}_t)$, with the convex output gate of Proposition 6 applied at readout. Best on MQAR (the state is purely accumulative); fails on overwrite (no mechanism to discard a stale binding).

GatedKATA (multiplicative scalar state gate). $\boldsymbol{S}_t = \gamma_t \boldsymbol{S}_{t-1} + \psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top$ with a learned scalar $\gamma_t \in [0, 1]$. The parameter-free convex output gate of Proposition 6 acts at readout and leaves $\boldsymbol{S}_t$ unchanged. The learned γ_t enters the recurrence and changes the relative weights stored in the state. Unrolling assigns write s the multiplicative weight $\prod_{r=s+1}^{t} \gamma_r$, producing token-dependent positional weighting that generally favors recent writes. The overwrite gains identify this positional weighting as the gate's leading role, with state-norm control as a secondary effect; see Section 4.2.

GatedDeltaKATA (PSD inside the GDN raw-key recurrence). The recurrence is:

$$\boldsymbol{S}_t = \gamma_t \big[\boldsymbol{S}_{t-1} - \beta_t \psi(\boldsymbol{k}_t)\psi(\boldsymbol{k}_t)^\top \boldsymbol{S}_{t-1}\big] + \beta_t \psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top.$$

This is Gated DeltaNet's recurrence with the rank-one erase term applied to $\psi(\boldsymbol{k}_t)$; GDN applies the same term to the raw $\boldsymbol{k}_t$. The construction inherits GDN's overwrite semantics and uses a Welch-cone PSD feature as its key substrate.

Comparison with GDN's raw-key geometry. Gated DeltaNet's recurrence treats the unit-normalized key $\boldsymbol{k}_t \in \mathbb{S}^{d-1}$ as the feature itself. Under the nonnegative Lorentz lift, any strict interference tolerance $\epsilon < 1/2$ forces negative raw-key inner products and hence permits at most $\min\{d+1, 1+1/(1-2\epsilon)\}$ keys by Theorem 7. At the boundary $\epsilon = 1/2$, nonpositive inner products permit up to $2d$ keys; for $d_{\text{head}} = 128$, this boundary is 256 tokens, consistent with the empirical degradation between $K = 256$ and $K = 512$. The rank-one factors used by KATA-Mg live in $\mathbb{S}^p_+$ with $p = d_{\text{head}}/g$; KATA-Σg sums the corresponding block outer products within the same cone. Their packing is governed by Theorem 2: above the $1/p$ Welch scale, the spherical-cap construction gives exponential address capacity in p at any fixed interference tolerance. This larger geometric packing budget is consistent with the MQAR trend at comparable fixed-state budgets.

E.2 OUTPUT GATING, RECURRENCE GATING, AND CONTENT REPLACEMENT

Positional recency in repeated-key overwrite. RoPE resolves the ambiguity in-distribution by assigning the two writes different position-dependent features. Because this benchmark always targets the latest write, the model learns to favor the most recent association. In a general retrieval problem, the earlier association may remain correct, and this learned recency preference can select the newer value. The positional shortcut also extrapolates poorly in Table 6, where both softmax+RoPE and KATA-M2+RoPE deteriorate rapidly beyond the training distribution.

The convex output gate of Proposition 6 acts at readout and leaves the relative contributions stored in $\boldsymbol{S}_t$ unchanged. GatedKATA places its learned scalar inside the recurrence. Unrolling $\boldsymbol{S}_t = \gamma_t \boldsymbol{S}_{t-1} + \psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top$ gives:

$$\boldsymbol{S}_t = \sum_{s=1}^{t} \left(\prod_{r=s+1}^{t} \gamma_r \right) \psi(\boldsymbol{k}_s)\boldsymbol{v}_s^\top.$$

The recurrent gate assigns every write a multiplicative, token-dependent positional weight. At K=64, this weighting raises overwrite accuracy from 0.521 for KATA-M2 to 0.991 for GatedKATA-M2. The result supports a positional interpretation of the scalar gate, while state magnitude control plays a secondary role. Since the product usually shrinks with age, the gate favors recent writes.

The delta rule corrects the value stored at the current content address. For a unit-norm feature, let $\widehat{\boldsymbol{v}}_t = \boldsymbol{S}_{t-1}^\top \psi(\boldsymbol{k}_t)$ be the value currently retrieved at $\boldsymbol{k}_t$. The update is:

$$\boldsymbol{S}_t = \boldsymbol{S}_{t-1} + \beta_t \psi(\boldsymbol{k}_t)\big(\boldsymbol{v}_t - \widehat{\boldsymbol{v}}_t\big)^\top.$$

When $\beta_t = 1$, reading the same address after the update returns $\boldsymbol{v}_t$ exactly. When the earlier association should be preserved, a learned β_t can suppress the correction. The update strength therefore controls content-dependent replacement at each write.

F Experimental protocols and architectures

F.1 MQAR worked example and protocol

Each MQAR example packs K distinct key–value pairs and K queries into a single sequence of length $T = 4K$: the first $2K$ tokens interleave the K bindings $(k_1, v_1, \ldots, k_K, v_K)$, then the remaining $2K$ positions reissue the K keys in random order, each followed by an answer slot the model must fill. Loss is computed only at answer positions, so the model must (i) *store* all K bindings presented in the first half and (ii) *retrieve* the correct value an arbitrary number of steps later. Keys and values are drawn from $V = 8192$, so random guessing floors at $1/8192 \approx 1.2 \times 10^{-4}$. A worked K=4, T=16 example: At K=1024, T=4096: 1024 disjoint bindings must survive through

Table 9: **MQAR worked example with $K = 4$ and $T = 16$.**

position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
input	A	7	B	3	C	9	D	1	C	?	A	?	D	?	B	?
target										**9**		**7**		**1**		**3**

up to 4096 intervening tokens before the recurrent state is queried, with the state discriminating among 1024 possibilities at each answer slot.

F.2 Additive and convex output-gate ablation

At d_{model}=128 with one head and KATA-Σ2 features:

Table 10: **Additive and convex output-gate ablation.**

Gate	state (B)	T=1024	T=2048	T=3072	T=4096
Convex	1,066,496	0.997	0.864	0.567	0.335
Additive (GatedKATA-Σ4)	271,872	**0.999**	**0.955**	**0.782**	**0.572**

F.3 MQAR architecture and training

Every model uses the canonical Zoology `TransformerBlock`. Its state mixer is `Identity`, and its sequence mixer is `Hybrid(BaseConv, mixer)`, so one short depthwise convolution precedes every sequence operator. We use one head, following Zoology's convention $d_{\text{head}} = d_{\text{model}}$, no positional embeddings, a tied LM head, and vocabulary size $V = 8192$. Training uses batch size 256, AdamW with $\text{wd} = 0.1$, cosine annealing over 32 epochs, and a three-point learning-rate sweep $\{10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$ with the best run selected per row. Training mixes streams up to $K_{\text{train}} = 64$ at $T = 256$; evaluation uses $K \in \{128, 256, 512, 768, 1024\}$ and $T \in \{512, 1024, 2048, 3072, 4096\}$.

GatedDeltaKATA in the overwrite table uses the full GDN-style configuration: ShortConv(k=4) on $\boldsymbol{q}/\boldsymbol{k}/\boldsymbol{v}$, GDN-style decay parameterization $(A_{\log}, d_t\text{bias}, a_{\text{proj}})$, RMSNorm on the output, and v-expansion 2×. GatedDeltaKATA applies the rank-1 erase to $\psi(\boldsymbol{k}_t)$, while GDN applies it to the raw $\boldsymbol{k}_t$; this is the single functional difference. Total state 141 KB matches GDN's 133 KB; parameter counts match within 1%.

F.4 Per-method architectures and reproducibility

We list, for each row in Sections 4.1 and 4.2, the precise projections, gates, normalizations, and feature maps so that each architecture can be reproduced from a fresh `TransformerBlock` skeleton. All variants share the canonical Zoology outer loop described above (two-layer hybrid block,

BaseConv(k=3), no positional embeddings unless explicitly stated, tied LM head, vocab V=8192). Differences below concern only the sequence-mixer module. Throughout, $d \equiv d_{\text{model}}$; with $H = 1$, we have $d_{\text{head}} = d$.

Plain KATA (additive recurrence with denominator). Three projections $W_q, W_k, W_v \in \mathbb{R}^{d\times d}$ are followed by a reduced PSD feature map. KATA-$\Sigma 2$ splits $\boldsymbol{k}$ into $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{d/2}$ and uses $\psi(\boldsymbol{k}) = \epsilon \boldsymbol{I} + \boldsymbol{a}\boldsymbol{a}^\top + \boldsymbol{b}\boldsymbol{b}^\top$. KATA-$\Sigma 4$ splits it into $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_4 \in \mathbb{R}^{d/4}$ and uses $\psi(\boldsymbol{k}) = \epsilon \boldsymbol{I} + \frac{1}{4}\sum_{i=1}^{4} \boldsymbol{a}_i \boldsymbol{a}_i^\top$. The factor $1/4$ is the numerical scale used in the reported $\Sigma 4$ configuration. It cancels from normalized plain-KATA weights as $\epsilon \to 0$ and is retained as part of the unnormalized gated configuration. In either case, $n_\psi = m(m+1)/2$ for block width $m = d/g$. The chunkwise recurrence computes the numerator and denominator with the packed feature map, and returns $\boldsymbol{o}_t = \psi(\boldsymbol{q}_t)^\top \boldsymbol{S}_t / \psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_t$ followed by $W_o \in \mathbb{R}^{d\times d}$. The *extras* configuration adds depthwise ShortConv($k = 4$), per-head RMSNorm, and value expansion $2\times$; $\epsilon = 10^{-4}$.

KATA-Orthant. Same W_q, W_k, W_v projections; feature map $\psi(\boldsymbol{x}) = \mathrm{ReLU}(\boldsymbol{x}) + \epsilon \mathbf{1}$, post-feature dimension $n_\psi = d$. The recurrence is normalized linear attention, $\boldsymbol{S}_t = \boldsymbol{S}_{t-1} + \psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top$, $\boldsymbol{Z}_t = \boldsymbol{Z}_{t-1} + \psi(\boldsymbol{k}_t)$, $\boldsymbol{o}_t = (\psi(\boldsymbol{q}_t)^\top \boldsymbol{S}_t)/(\psi(\boldsymbol{q}_t)^\top \boldsymbol{Z}_t)$, implemented via FLA's normalized linear-attention kernel. ShortConv(k=4) and output $\mathrm{RMSNorm}$ are enabled; $\epsilon = 10^{-6}$.

KATA-Lorentz. Identical to KATA-Orthant except for the feature map: with $\boldsymbol{x} \in \mathbb{R}^d$, write $\boldsymbol{y} = \boldsymbol{x}_{1:d-1}$ and $s = x_d$, then $\psi(\boldsymbol{x}) = (\boldsymbol{y},\ \|\boldsymbol{y}\|_2(1+s^2)) \in \mathbb{R}^d$. The image lies in the closed Lorentz cone $\{(\boldsymbol{y}, t) : t \geq \|\boldsymbol{y}\|_2\}$ so $\langle \psi(\boldsymbol{x}), \psi(\boldsymbol{y})\rangle \geq 0$. State and dispatch are identical to the orthant case.

GatedKATA (additive scalar decay, no denominator). Adds a single per-head per-token log-decay projection $W_\alpha \in \mathbb{R}^{d\times H}$ (with bias) producing $\log \gamma_t = \log \sigma(W_\alpha \boldsymbol{x}_t) \in \mathbb{R}^H$. The recurrence is $\boldsymbol{S}_t = \gamma_t \boldsymbol{S}_{t-1} + \psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top$, $\boldsymbol{o}_t = \psi(\boldsymbol{q}_t)^\top \boldsymbol{S}_t$ (no denominator), implemented via FLA's simple gated-linear-attention kernel on the packed ψ with the per-token gate. The *convex-gate* ablation pre-scales $\boldsymbol{v}_t$ by $1 - \gamma_t$ before the same kernel, yielding $\boldsymbol{S}_t = \gamma_t \boldsymbol{S}_{t-1} + (1-\gamma_t)\psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top$. The KATA-$\Sigma 2$ and KATA-$\Sigma 4$ feature variants are as above; $\epsilon = 10^{-4}$.

GatedDeltaKATA (PSD inside the GDN raw-key recurrence). Component-for-component match to Gated DeltaNet except for the substitution of $\psi(\boldsymbol{k}_t)$ for $\boldsymbol{k}_t$ in the recurrence. Projections: $W_q, W_k \in \mathbb{R}^{d\times d}$, $W_v \in \mathbb{R}^{d\times 2d}$ (default v-expansion $2\times$). Depthwise ShortConv(k=4) on $\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}$, no SiLU. Beta projection $W_\beta \in \mathbb{R}^{d\times H}$ producing $\beta_t = \sigma(W_\beta \boldsymbol{x}_t)$. Decay parameterization, matching FLA's gated-delta-rule kernel, $g_t = -\exp(A_{\log}) \odot \mathrm{softplus}(W_\alpha \boldsymbol{x}_t + \mathrm{dt_bias})$ with $A_{\log} \in \mathbb{R}^H$ initialized by $\log U(0, 16)$ and $\mathrm{dt_bias} \in \mathbb{R}^H$ calibrated so that $\mathrm{softplus}(\mathrm{dt_bias}) \sim U(10^{-3}, 10^{-1})$; $A_{\log}$ and $\mathrm{dt_bias}$ are excluded from weight decay. The recurrence is:

$$\boldsymbol{S}_t = \gamma_t \big[\boldsymbol{S}_{t-1} - \beta_t \psi(\boldsymbol{k}_t)\psi(\boldsymbol{k}_t)^\top \boldsymbol{S}_{t-1}\big] + \beta_t \psi(\boldsymbol{k}_t)\boldsymbol{v}_t^\top.$$

This recurrence is delegated to FLA's gated-delta-rule kernel on packed ψ; the rank-1 erase $\psi(\boldsymbol{k}_t)^\top \boldsymbol{S}_{t-1}$ is contracted directly, so the $n_\psi \times n_\psi$ matrix $\psi(\boldsymbol{k}_t)\psi(\boldsymbol{k}_t)^\top$ is never materialized. Output passes through per-head $\mathrm{RMSNorm}$ ($\epsilon_{\text{norm}} = 10^{-5}$) and $W_o \in \mathbb{R}^{2d\times d}$. KATA-$\Sigma 4$ uses $H = 2$ because the kernel requires packed feature dimension at most 256; the packed dimension is 528 at $H = 1, d = 128$ and 136 at $H = 2$. Optional knobs (disabled in the reported runs unless noted): per-head ℓ_2 normalization of $\boldsymbol{q}, \boldsymbol{k}$ before ψ with $\epsilon = 10^{-6}$; RoPE on $\boldsymbol{q}, \boldsymbol{k}$ with base 10^4.

Gated DeltaNet (baseline). FLA's `GatedDeltaNet` mixer used out-of-the-box, configured to match the surrounding skeleton: `l_max` set to the longest evaluation sequence, NH= 1, `use_short_conv` =`True` (k=4), `use_gate` =`False`. Feature substrate: raw $\boldsymbol{k}_t \in \mathbb{S}^{d-1}$ (FLA's `use_qk_l2norm_in_kernel`). Recurrence is identical in form to GatedDeltaKATA above with $\psi(\boldsymbol{k}) = \boldsymbol{k}$. State per head is $d \times 2d$ (for v-expansion 2); at d=128, NH= 1, total $\sim$133 KB.

Softmax attention (full / sliding window). A standard MHA block uses PyTorch scaled dot-product attention (SDPA) with a causal mask. Three projections $W_q, W_k, W_v \in \mathbb{R}^{d\times d}$ and an output $W_o \in \mathbb{R}^{d\times d}$ surround the attention operator. The +RoPE variant rotates $\boldsymbol{q}, \boldsymbol{k}$ before SDPA with base $\theta \in \{10^4, 5\times 10^4\}$; reported runs use $\theta = 5\times 10^4$. The sliding-window variant restricts the mask to

the $W = 256$ tokens preceding the query and caches the boolean mask. We use PyTorch's memory-efficient SDPA backend for this configuration. At sequence length T, the KV cache contains $2dT$ floats for full softmax and $2dW$ for sliding-window attention.

Initialization and optimization. Linear layers use Zoology's $\mathcal{N}(0, 0.02^2)$ initialization, with GPT-2 residual scaling $1/\sqrt{2N_{\text{layers}}}$ on W_o, and the LM head is tied to the embeddings. We use AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$), weight decay 0.1 except on $A_{\log}$ and $\mathrm{dt_bias}$, gradient clipping at 1.0, and a 32-epoch cosine schedule without warmup. For each row, we select the best learning rate in $\{10^{-3}, 3\times10^{-3}, 10^{-2}\}$ and stop when the hardest extrapolation slice reaches 0.999 accuracy.

G FUTURE DIRECTIONS

Future work. Four directions follow directly. First, packing theory and constructive dictionaries should be extended to complex spaces, where spherical codes and ETFs may offer more capacity at equal dimension. Second, GatedKATA and GatedDeltaKATA should be pretrained systematically, while DeltaKATA should be evaluated beyond M1; all three require dedicated fused GPU kernels. Third, the learnability of the full KATA-M1 geometry should be tested beyond $d_{\text{head}} = 64$, together with kernels that make its quadratic feature state practical at wider heads. Finally, tensor-train factorizations (Oseledets, 2011) offer a path to higher even-order outer-product features without materializing the full tensor state. These studies will determine whether the geometric gains persist as feature order, model scale, and contextual demands increase.

H HARDWARE-ALIGNED IMPLEMENTATION

We implement KATA as fused Triton kernels at two operating points: a flash-attention-style forward up to $\sim$1.6$\times$ FlashAttention-2 throughput, and an exact $O(T)$ chunked-state form that reaches $\sim$11$\times$ FlashAttention-2 forward throughput at 131k tokens. An associative scan of the first-order feature lowers the inter-chunk recurrence depth to $O(\log(T/C))$ for chunk size C and averages $\sim$2.4$\times$ the throughput of a matched sequential linear-attention baseline.

The capacity story of Section 3 is geometric, but its practical reach depends on how cleanly the feature map maps onto modern accelerators. Chunkwise parallel training is the standard scaffold for linear-time recurrent attention: split a length-T sequence into chunks of size C, yielding $N_C := \lceil T/C \rceil$ chunks with the final chunk padded when necessary, run intra-chunk attention on tensor cores, and propagate the inter-chunk state by a sequential prefix sum (Hua et al., 2022). FLA's chunk-parallel linear attention (Yang & Zhang, 2024) and Gated DeltaNet's non-commutative Householder transition in the compact WY representation (Schreiber & Van Loan, 1989; Yang et al., 2025b) share this decomposition; what differs is the inter-chunk update. GDN requires a $C\times C$ triangular solve per chunk, expanding the per-token state into high-bandwidth memory (HBM) and tying chunk size to the solve's conditioning. Mamba2's data-dependent transition is likewise sequential at the chunk level. An additive feature-state recurrence avoids both costs and admits a parallel scan that lowers inter-chunk depth from $\Theta(N_C)$ to $\mathcal{O}(\log N_C)$.

H.1 FROM RANK-ONE RAYS TO LOW-RANK PSD PACKING

From this point onward, unqualified references to a *KATA mixer* or *KATA kernel* denote PSD-resident features. Orthant and Lorentz variants are named explicitly. The full construction uses the rank-one map $\psi(\boldsymbol{u}) = \boldsymbol{u}\boldsymbol{u}^\top$. A direct outer-product layout has d^2 entries for head dimension d, or $n = d(d+1)/2$ unique coordinates after symmetry packing. Thus $n_\psi = n$ for the full PSD map. Linear attention must retain the feature–value state:

$$\boldsymbol{S}_t = \Psi(\boldsymbol{K}_{1:t})^\top \boldsymbol{V}_{1:t} \in \mathbb{R}^{n\times d_v},$$

so its dominant storage is nd_v; the feature width alone understates the memory cost. At $d = 64$, the direct and packed widths are already 4096 and 2080, respectively. Multiplication by d_v places the full PSD map at the edge of hardware-friendly state sizes even at this standard head dimension.

We therefore test two reduced PSD families. Let $g \in \mathbb{N}$ divide d, set $m = d/g$, and partition each projected query or key into g blocks $\boldsymbol{u} = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_g)$ with $\boldsymbol{u}_a \in \mathbb{R}^m$. KATA-$\Sigma g$ sums the block outer products into one PSD feature:

$$\psi_{\Sigma g}(\boldsymbol{u}) = \sum_{a=1}^{g} \boldsymbol{u}_a \boldsymbol{u}_a^\top \in \mathbb{S}_+^m, \qquad \langle \psi_{\Sigma g}(\boldsymbol{q}), \psi_{\Sigma g}(\boldsymbol{k}) \rangle_F = \sum_{a,b=1}^{g} (\boldsymbol{q}_a^\top \boldsymbol{k}_b)^2.$$

Its dominant state size is:

$$\frac{m(m+1)}{2} d_v = \frac{(d/g)(d/g+1)}{2} d_v.$$

Thus KATA-Σg compresses g outer products into a single low-rank PSD factor.

KATA-Mg instead retains one PSD factor per block. The raw inner product decomposes as $\boldsymbol{q}^\top \boldsymbol{k} = \sum_{a=1}^{g} \boldsymbol{q}_a^\top \boldsymbol{k}_a$, while the blockwise PSD map is:

$$\psi_{\mathrm{M}g}(\boldsymbol{u}) = \left(\boldsymbol{u}_1\boldsymbol{u}_1^\top, \dots, \boldsymbol{u}_g\boldsymbol{u}_g^\top\right) \in (\mathbb{S}_+^m)^g, \qquad \langle \psi_{\mathrm{M}g}(\boldsymbol{q}), \psi_{\mathrm{M}g}(\boldsymbol{k}) \rangle = \sum_{a=1}^{g} (\boldsymbol{q}_a^\top \boldsymbol{k}_a)^2.$$

This maps into the Cartesian product of g copies of $\mathbb{S}_+^{d/g}$ and has dominant state size:

$$g\frac{m(m+1)}{2} d_v = \frac{d(d/g+1)}{2} d_v.$$

Both families recover the full rank-one PSD map at $g = 1$. A *Delta* prefix adds the delta rule, a *Gated* prefix adds the learned gate, and *GatedDelta* combines both without changing the underlying PSD feature geometry. Their recurrences and implementation details are specified in Appendices E, F.4 and H.3.

Chunkwise forward. Let n_ψ denote the active feature dimension. We partition the sequence into chunks of size C and write the per-chunk feature matrices as $\Psi(\boldsymbol{Q}_{[t]}), \Psi(\boldsymbol{K}_{[t]}) \in \mathbb{R}^{C\times n_\psi}$ and $\boldsymbol{V}_{[t]} \in \mathbb{R}^{C\times d_v}$. The inter-chunk state update reads:

$$\boldsymbol{S}_{[t+1]} = \boldsymbol{S}_{[t]} + \Psi(\boldsymbol{K}_{[t]})^\top \boldsymbol{V}_{[t]}, \quad \boldsymbol{Z}_{[t+1]} = \boldsymbol{Z}_{[t]} + \Psi(\boldsymbol{K}_{[t]})^\top \mathbf{1}_C. \tag{24}$$

With $\boldsymbol{M} \in \{0,1\}^{C\times C}$ the lower-triangular causal mask and $\boldsymbol{P} = (\Psi(\boldsymbol{Q}_{[t]})\Psi(\boldsymbol{K}_{[t]})^\top) \odot \boldsymbol{M}$, the chunk output is:

$$\boldsymbol{N}_{[t]} = \Psi(\boldsymbol{Q}_{[t]})\boldsymbol{S}_{[t]} + \boldsymbol{P}\,\boldsymbol{V}_{[t]} \tag{25}$$

$$\boldsymbol{D}_{[t]} = \Psi(\boldsymbol{Q}_{[t]})\boldsymbol{Z}_{[t]} + \boldsymbol{P}\,\mathbf{1}_C \tag{26}$$

$$\boldsymbol{O}_{[t]} = \mathrm{diag}(\boldsymbol{D}_{[t]})^{-1}\boldsymbol{N}_{[t]} \tag{27}$$

Every operation in this pass is a dense general matrix multiplication (GEMM), an element-wise op, or a warp reduction. Relative to Gated DeltaNet's WY representation, which forms and inverts a $C\times C$ unitriangular matrix per chunk, the inter-chunk cost drops from $\mathcal{O}(Cd_k^2+d_k^2d_v)$ to $\mathcal{O}(Cn_\psi d_v)$, and available SRAM sets the chunk-size limit. Triangular-solve conditioning no longer constrains it. The full pass is summarized in Algorithm 1.

H.2 Chunkwise forward pass

Algorithm 1 KATA chunkwise forward pass for one head and one layer

Require: $\boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{T\times d}$, $\boldsymbol{V} \in \mathbb{R}^{T\times d_v}$, chunk size C, cone map ψ

1: $N_C \leftarrow \lceil T/C \rceil$; pad the final chunk to length C
2: $\boldsymbol{S}_{[0]} \leftarrow \mathbf{0}_{n_\psi\times d_v}$, $\boldsymbol{Z}_{[0]} \leftarrow \mathbf{0}_{n_\psi}$
3: **for** $t = 0, \ldots, N_C - 1$ **do**
4: load $\boldsymbol{Q}_{[t]}, \boldsymbol{K}_{[t]}, \boldsymbol{V}_{[t]}$ into SRAM
5: $\Psi(\boldsymbol{Q}_{[t]}), \Psi(\boldsymbol{K}_{[t]}) \leftarrow \psi(\boldsymbol{Q}_{[t]}), \psi(\boldsymbol{K}_{[t]})$ ▷ in-SRAM expansion
6: compute $\boldsymbol{O}_{[t]}$ via Equation (27)
7: update $\boldsymbol{S}_{[t+1]}, \boldsymbol{Z}_{[t+1]}$ via Equation (24)
8: **end for**
9: **return** $\boldsymbol{O} = [\boldsymbol{O}_{[0]}; \ldots; \boldsymbol{O}_{[N_C-1]}]$, truncated to T tokens

H.3 Delta-rule chunked forward and backward

Recurrence. The delta variant replaces the additive write with a content-based *erase* applied in feature space. With per-token strength $\beta_t \in (0,1)$, the pseudo-values $\boldsymbol{u}_t \in \mathbb{R}^{d_v}$ are defined by forward substitution on a lower-triangular system, and the readout reads them out through the PSD kernel:

$$\boldsymbol{u}_t = \beta_t\Big(\boldsymbol{v}_t - \sum_{s<t}\langle\psi(\boldsymbol{k}_s), \psi(\boldsymbol{k}_t)\rangle\,\boldsymbol{u}_s\Big), \qquad \boldsymbol{o}_t = \sum_{s\le t}\langle\psi(\boldsymbol{q}_t), \psi(\boldsymbol{k}_s)\rangle\,\boldsymbol{u}_s. \tag{28}$$

The erase kernel is the PSD Gram $\langle\psi(\boldsymbol{k}_s), \psi(\boldsymbol{k}_t)\rangle = \sum_{ij}(\boldsymbol{k}_{s,i}\cdot\boldsymbol{k}_{t,j})^2$ (KATA-Σg, with i,j over the g groups). GDN uses the raw $\boldsymbol{k}_s\cdot\boldsymbol{k}_t$ kernel. The readout is *unnormalized* (no denominator), and the online Widrow–Hoff erase keeps $\boldsymbol{S} = \sum_s \psi(\boldsymbol{k}_s)\boldsymbol{u}_s^\top \in \mathbb{R}^{n_\psi\times d_v}$ bounded.

Chunked forward (WY-free). Let $\boldsymbol{S}$ be the state entering chunk $[c]$. Form the two $C\times C$ Gram blocks $A^{kk}_{ts} = \sum_{ij}(\boldsymbol{k}_{t,i}\cdot\boldsymbol{k}_{s,j})^2$ and $A^{qk}_{ts} = \sum_{ij}(\boldsymbol{q}_{t,i}\cdot\boldsymbol{k}_{s,j})^2$ *directly from the raw projections*; no $n_\psi\times n_\psi$ erase matrix is formed. With $\boldsymbol{N} = \mathrm{diag}(\boldsymbol{\beta})\,\mathrm{tril}(A^{kk}, -1)$:

$$\boldsymbol{V}' = \boldsymbol{V}_{[c]} - \Psi(\boldsymbol{K}_{[c]})\,\boldsymbol{S}, \qquad \boldsymbol{T} = (\boldsymbol{I}+\boldsymbol{N})^{-1} = \textstyle\sum_{r\ge 0}(-\boldsymbol{N})^r, \tag{29}$$

$$\boldsymbol{U} = \boldsymbol{T}\,\mathrm{diag}(\boldsymbol{\beta})\,\boldsymbol{V}', \qquad \boldsymbol{O}_{[c]} = \Psi(\boldsymbol{Q}_{[c]})\,\boldsymbol{S} + \mathrm{tril}(A^{qk})\,\boldsymbol{U}, \tag{30}$$

$$\boldsymbol{S}' = \boldsymbol{S} + \Psi(\boldsymbol{K}_{[c]})^\top\boldsymbol{U}. \tag{31}$$

Since $\boldsymbol{N}$ is strictly lower-triangular (nilpotent, $\boldsymbol{N}^C{=}0$), the inverse $\boldsymbol{T}$ is exact after $\lceil\log_2 C\rceil$ Neumann doublings; only $C\times C$, $C\times d_v$, and the $n_\psi\times d_v$ state ever reside in SRAM.

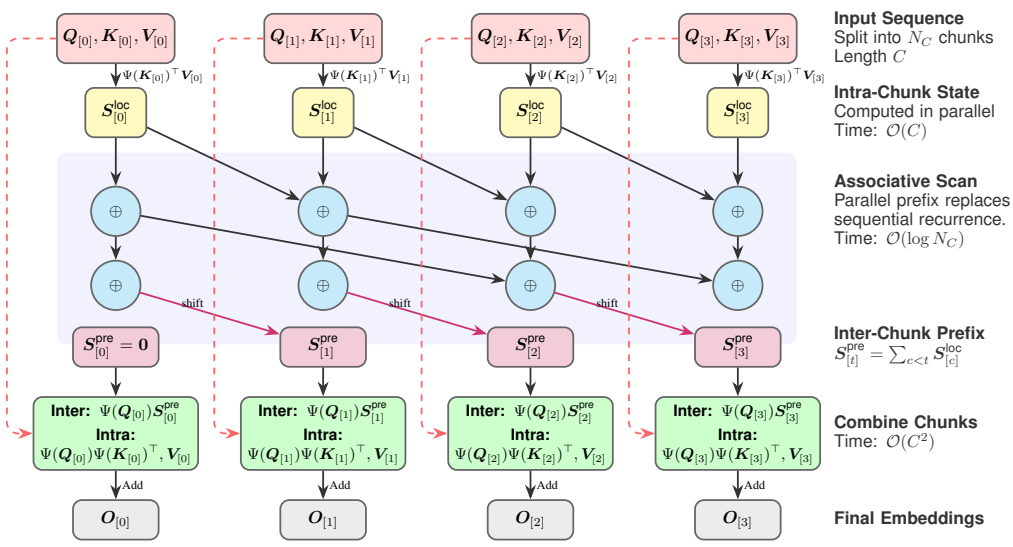


Figure 2: **Chunkwise associative scan.** Each chunk forms its local state in parallel. An exclusive scan produces inter-chunk prefixes in $\lceil \log_2 N_C \rceil$ rounds, and each output kernel combines its prefix with the local $C \times C$ causal term. Only $\boldsymbol{S}$ is shown; $\boldsymbol{Z}$ follows the same scan.

Chunked backward. The backward is a reverse scan over chunks carrying the state cotangent $\mathrm{d}\boldsymbol{S}_+$ (from later chunks); the forward quantities $\boldsymbol{S}, \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V}'$ are recomputed from the checkpointed $\boldsymbol{S}$. Per chunk (all $+=$ accumulate into $\mathrm{d}\boldsymbol{S}$, returned to chunk $[c-1]$):

$$
\begin{aligned}
\mathrm{d}\Psi(\boldsymbol{Q}) &= \mathrm{d}\boldsymbol{O}_{[c]}\boldsymbol{S}^\top, \\
\mathrm{d}\boldsymbol{S} &= \Psi(\boldsymbol{Q})^\top \mathrm{d}\boldsymbol{O}_{[c]} + \mathrm{d}\boldsymbol{S}_+, \\
\mathrm{d}A^{qk} &= \mathrm{tril}\big(\mathrm{d}\boldsymbol{O}_{[c]}\boldsymbol{U}^\top\big), \\
\mathrm{d}\boldsymbol{U} &= \mathrm{tril}(A^{qk})^\top \mathrm{d}\boldsymbol{O}_{[c]} + \Psi(\boldsymbol{K})\mathrm{d}\boldsymbol{S}_+, \\
\mathrm{d}\boldsymbol{T} &= \mathrm{d}\boldsymbol{U}\big(\mathrm{diag}(\boldsymbol{\beta})\boldsymbol{V}'\big)^\top, \qquad \mathrm{d}\boldsymbol{V}' = \mathrm{diag}(\boldsymbol{\beta})\boldsymbol{T}^\top \mathrm{d}\boldsymbol{U}, \\
\mathrm{d}\boldsymbol{N} &= -\boldsymbol{T}^\top \mathrm{d}\boldsymbol{T}\,\boldsymbol{T}^\top, \qquad \mathrm{d}A^{kk} = \mathrm{diag}(\boldsymbol{\beta})\,\mathrm{tril}(\mathrm{d}\boldsymbol{N}, -1), \\
\mathrm{d}\boldsymbol{V}_{[c]} &= \mathrm{d}\boldsymbol{V}', \qquad \mathrm{d}\boldsymbol{S} \mathrel{-}= \Psi(\boldsymbol{K})^\top \mathrm{d}\boldsymbol{V}', \\
\mathrm{d}\Psi(\boldsymbol{K}) &= \boldsymbol{U}\mathrm{d}\boldsymbol{S}_+^\top - \mathrm{d}\boldsymbol{V}'\boldsymbol{S}^\top \\
&\quad + (\mathrm{d}A^{kk} + \mathrm{d}A^{kk\top})\Psi(\boldsymbol{K}) + \mathrm{d}A^{qk\top}\Psi(\boldsymbol{Q}), \\
\mathrm{d}\Psi(\boldsymbol{Q}) &\mathrel{+}= \mathrm{d}A^{qk}\Psi(\boldsymbol{K}), \\
\mathrm{d}\boldsymbol{\beta} &= \mathrm{rowsum}\big(\boldsymbol{V}' \odot \boldsymbol{T}^\top \mathrm{d}\boldsymbol{U}\big) \\
&\quad + \mathrm{rowsum}\big(\mathrm{tril}(A^{kk}, -1) \odot \mathrm{d}\boldsymbol{N}\big).
\end{aligned}
\tag{32}
$$

The only non-matmul VJP is the matrix-inverse term $\mathrm{d}\boldsymbol{N} = -\boldsymbol{T}^\top \mathrm{d}\boldsymbol{T}\,\boldsymbol{T}^\top$; the feature-map VJP closes the chain, $\mathrm{d}\boldsymbol{k}_g = (\mathrm{d}\Psi(\boldsymbol{K})_g + \mathrm{d}\Psi(\boldsymbol{K})_g^\top)\boldsymbol{k}_g$ per group (and likewise for $\boldsymbol{q}$). The backward therefore has the same $\mathcal{O}\big(TC + T\,n_\psi\,d_v/C\big)$ cost profile as the forward, with a single $n_\psi \times d_v$ state cotangent streamed across chunks.

H.4 Tree-scan inter-chunk reduction

A useful observation is that the recurrence Eq. (24) is an associative prefix sum across the N_C chunks. Existing chunk-parallel implementations, including the FLA and Gated DeltaNet chunk paths, traverse this prefix sequentially with depth $\Theta(N_C)$. Because $(\boldsymbol{S}_{\mathrm{loc}}, \boldsymbol{Z}_{\mathrm{loc}})$ is a vector-space cumulative sum, however, the same prefix admits a parallel scan with depth $\mathcal{O}(\log_2 N_C)$. Our forward decomposes into three Triton kernels dispatched in series: a per-chunk *state* kernel with $B \cdot H \cdot N_C$ programs in parallel, a *cumulative* pass that performs a associative scan in $\lceil \log_2 N_C \rceil$

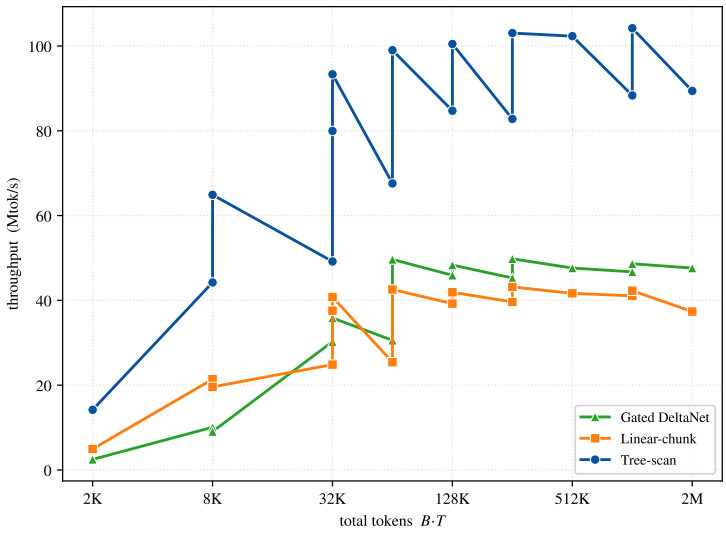


Figure 3: **Forward throughput over $B \times T$.** Associative scan, sequential scan, and Gated DeltaNet on H100.

rounds with N_C-wide parallelism per round, and a per-chunk *output* kernel that combines the inter-chunk prefix with the local $C \times C$ causal pattern. The resulting wall-clock model is:

$$t_{\text{total}} = \underbrace{\mathcal{O}\big(Cn_\psi d_v\big)}_{\text{state}} + \underbrace{\mathcal{O}\big(\log_2 N_C \cdot d_v\big)}_{\text{tree-scan}} + \underbrace{\mathcal{O}\big(Cn_\psi d_v + C^2(n_\psi + d_v)\big)}_{\text{output}},$$

replacing the linear-scan baseline's $\mathcal{O}(N_C d_v)$ middle term. The n_ψ feature slices are independent and execute in parallel. At T=2048 and C=32, the scan depth shrinks from N_C=64 to $\log_2 N_C$=6. Its arithmetic cost is negligible relative to the HBM write of $\boldsymbol{S}_{\text{loc}}$ at chunk boundaries, which is the actual bottleneck.

Occupancy at small batch. The advantage is largest precisely where a sequential scan starves the GPU. At B=1, the $B \cdot H$=16 programs of a sequential chunked scan, as in FLA or Gated DeltaNet, cannot fill the H100's 132 SMs, so the forward is occupancy-bound and plateaus near 25 Mtok/s. The tree scan instead exposes $B \cdot H \cdot N_C$ programs, which scale with the actual work. Figure 3 and Table 17 report the full $B \times T$ grid and its $B = 1$ slice, respectively. For the linear feature ($\psi(\boldsymbol{x}) = \boldsymbol{x}$, with a matched $d \times d$ state), the associative scan delivers $\mathbf{1.9}$–$\mathbf{2.9}\times$ the throughput of the sequential linear-chunk kernel from 2048 to 131k tokens, and up to $7.5\times$ the throughput of Gated DeltaNet at $T = 2048$.

Why the PSD feature keeps to $\mathcal{O}(T)$ or $\mathcal{O}(T^2)$. The tree scan is viable only for the linear feature. A PSD block of width $m = d/g$ has a direct outer-product feature dimension m^2, so a chunk's local state $\psi(\boldsymbol{K}_{[c]})^\top \boldsymbol{V}_{[c]}$ is an $m^2 \times d_v$ tile. It occupies 128 KB at g=2 (m^2=1024) and already exceeds the H100's 228 KB shared-memory budget once the query feature and matrix-multiplication accumulators are resident. At g=1 (m^2=4096), it occupies 1.5 MB. A scan must moreover materialize every chunk's state to HBM, requiring $\mathcal{O}(N_C m^2 d_v)$ storage traffic. We therefore do not tree-scan the PSD recurrence. KATA-Mg runs either as a quadratic $\mathcal{O}(T^2)$ kernel, which never forms the m^2 state, or as a linear-state $\mathcal{O}(T)$ kernel, whose $m^2 \times d_v$ state lives in HBM and is streamed in feature blocks. Realizing the parallel-prefix speedup for PSD features would require this feature blocking within the scan and is left to future work. Full timings and SRAM budgets are reported in Appendices H.4.1 and H.5.2.

H.4.1 Associative-scan kernel

The chunked attention forward decomposes into three stages, each launched as a separate Triton kernel that communicates with the next through small HBM buffers at chunk boundaries. Let C

denote the chunk size and define the number of chunks as:

$$N_C := \left\lceil \frac{T}{C} \right\rceil .$$

The final chunk is padded when C does not divide T. The analysis assumes already-feature-mapped inputs $\boldsymbol{q} = \psi(\hat{\boldsymbol{q}})$ and $\boldsymbol{k} = \psi(\hat{\boldsymbol{k}})$ in $\mathbb{R}^{n_\psi}$, where n_ψ denotes the active post-feature dimension.

Three-kernel pipeline. (i) *Chunk-state* computes, for every chunk c:

$$\boldsymbol{S}_{\text{loc}}^{(c)} = \boldsymbol{K}_{[c]}^\top \boldsymbol{V}_{[c]} \in \mathbb{R}^{n_\psi \times d_v}, \qquad \boldsymbol{Z}_{\text{loc}}^{(c)} = \mathbf{1}^\top \boldsymbol{K}_{[c]} \in \mathbb{R}^{n_\psi}.$$

All (b, h, c) triples are independent, exposing $B{\cdot}H{\cdot}N_C$ programs. (ii) *Inter-chunk reduction* computes the exclusive prefix:

$$(\boldsymbol{S}_{\text{pre}}^{(c)}, \boldsymbol{Z}_{\text{pre}}^{(c)}) = \sum_{c'<c} (\boldsymbol{S}_{\text{loc}}^{(c')}, \boldsymbol{Z}_{\text{loc}}^{(c')})$$

along the chunk axis. This is the only stage with a cross-chunk dependency. (iii) *Chunk-output* combines the inter-chunk prefix with the local $C \times C$ causal pattern to produce $\boldsymbol{O}_{[c]}$.

Associative scan versus sequential scan. The standard chunk-parallel implementation (FLA's linear-attention kernel (Yang & Zhang, 2024)) evaluates the prefix sequentially over N_C chunks, giving depth $\Theta(N_C)$. Our associative scan uses $\lceil \log_2 N_C \rceil$ rounds, each parallel across the N_C chunk slots. Its total work is $\mathcal{O}(N_C \log N_C \cdot n_\psi d_v)$, compared with $\mathcal{O}(N_C n_\psi d_v)$ for the linear scan, while its parallel depth is $\mathcal{O}(\log N_C \cdot d_v)$. With one program per batch index b, head index h, and feature index $i \in \{1, \ldots, n_\psi\}$, all $B{\cdot}H{\cdot}n_\psi$ feature slices scan independently.

Wall-clock model. The three stages have wall-clock cost:

$$t_{\text{total}} = \underbrace{\mathcal{O}(C n_\psi d_v)}_{\text{chunk-state}} + \underbrace{\mathcal{O}(\log N_C \cdot d_v)}_{\text{tree-scan}} + \underbrace{\mathcal{O}(C n_\psi d_v + C^2(n_\psi + d_v))}_{\text{chunk-output}} .$$

The middle term replaces the linear-scan baseline's $\mathcal{O}(N_C d_v)$. At $T{=}2048$ and $C{=}32$, the scan depth decreases from $N_C{=}64$ to $\log_2 N_C{=}6$. In practice, the scan stage is a small fraction of the wall-clock time. The dominant cost is the HBM write of $\boldsymbol{S}_{\text{loc}}$ and the subsequent prefix read. Reducing the per-chunk state through a smaller n_ψ, as in orthant, Lorentz, or grouped PSD features, therefore translates directly into higher throughput.

Measured throughput (H100 SXM, bf16, $B{=}4$, $H{=}12$, $d_v{=}64$). Throughput in million tokens/s, $C{=}32$, against FLA's linear-attention kernel on the same materialized inputs:

Table 11: **Measured associative-scan throughput.**

T	N_C	feature map	n_ψ	tree (Mtok/s)	linear (Mtok/s)	FLA (Mtok/s)
1024	32	orthant	64	**25.6**	24.1	10.78
1024	32	lorentz	64	**25.6**	24.1	12.05
2048	64	orthant	64	**30.3**	27.3	23.41
2048	64	lorentz	64	**30.3**	27.3	23.41
4096	128	orthant	64	33.5	29.8	**37.2**

At $T \leq 2048$, the tree-scan kernel delivers 2.1–2.4$\times$ the throughput of FLA at $T{=}1024$ and 1.3$\times$ at $T{=}2048$; tree versus linear within the same dispatch is a consistent 1.06–1.12$\times$ throughput gain that grows with N_C. At $T{=}4096$, chunk-state becomes HBM-bound, and FLA's internal n_ψ-axis tiling overtakes our single-tile output kernel. The asymptotic $\log N_C$ advantage is then hidden by the bandwidth ceiling on writes of $\boldsymbol{S}_{\text{loc}}$. For PSD-packed $n_\psi \geq 1024$, FLA's pre-materialize path with internal feature-axis tiling currently wins. Adding the same tiling to the tree-scan output kernel is the natural follow-on.

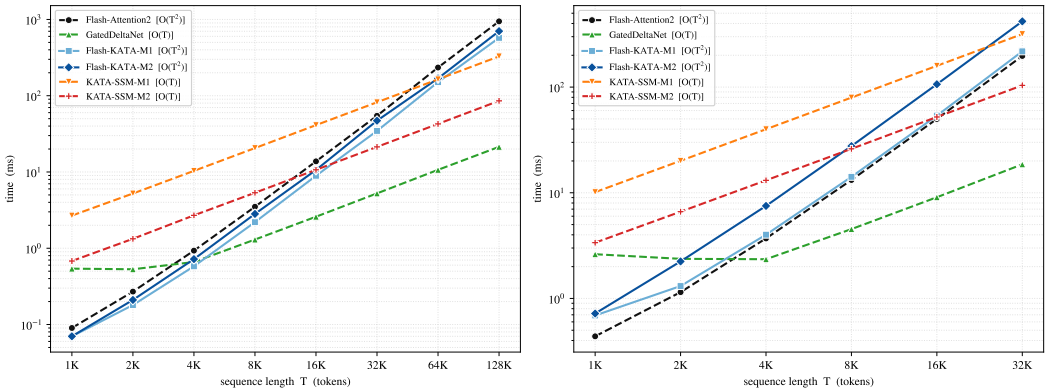


Figure 4: **KATA-Mg latency versus sequence length on an NVIDIA H100.** Forward latency (left) and forward-plus-backward latency (right) at B=8, H=16, d_{head}=64, and bf16 on log–log axes. Solid curves are the $\mathcal{O}(T^2)$ kernels, FlashAttention-2 and quadratic KATA-Mg; dashed curves are the $\mathcal{O}(T)$ recurrences, Gated DeltaNet and linear-state KATA-Mg. Lower is better; exact timings appear in Tables 14 and 15.

H.5 Synthetic induction and hardware microbenchmarks

Induction preserves near-perfect recall, while kernel benchmarks show that the maps can be implemented efficiently.

KATA-Σ2 solves a V=256 induction-head probe at every distance up to 2048 tokens within an 8k-step budget, matching or nearly matching softmax with FlashAttention-2 at every scale tested while avoiding the $\mathcal{O}(T^2)$ memory of dense attention (Appendix H.5.1). The fused rank-two KATA-Σ2 forward scales linearly with T at near-peak bandwidth without writing the per-token ψ tensor to HBM; full timings appear in Appendix H.5.2.

Kernel latency and throughput ratios. Figure 4 sweeps sequence length on an NVIDIA H100 for the two computational forms of the KATA-Mg kernel. Across the tested lengths, the quadratic KATA-M1 forward delivers 1.3–1.6$\times$ the throughput of FlashAttention-2, while KATA-M2 delivers 1.2–1.4$\times$ the throughput. Both retain the same $\mathcal{O}(T^2)$ scaling. The exact linear-state PSD kernel instead trades an $m^2 \times d_v$ recurrent state, with $m = d/g$, for $\mathcal{O}(T)$ compute. KATA-M2 reaches parity with its quadratic counterpart near 16K and is faster by 32K; KATA-M1 is faster by 128K. At 131K, linear-state KATA-M2 reaches 11$\times$ FlashAttention-2's forward throughput while implementing the exact PSD recurrence. Gated DeltaNet remains faster, but its raw-key feature loses high-entropy needle information (Table 2). The measured latency scaling approaches 2$\times$ per doubling for the linear kernels and 4$\times$ for the quadratic kernels.

H.5.1 Induction-head benchmark

Table 12: **Induction-head benchmark.**

	d=128	d=256	d=512	d=1024	d=2048
Softmax (FA) acc	1.00	1.00	1.00	1.00	1.00
Softmax (FA) phase step	2k	2k	3k	3k	4k
KATA-Σ2 acc	1.00	1.00	1.00	1.00	0.99
KATA-Σ2 phase step	2k	2k	2k	3k	8k

The two model families are aligned on the Zoology (Arora et al., 2023) backbone: two-layer pre-norm transformer, d_{model}=128, four heads, 4$\times$ GELU state mixer, LM head weight-tied to the token embedding, Zoology-style Linear/Embedding init with std= 0.02 and GPT-2 residual scaling. Softmax uses `F.scaled_dot_product_attention`; KATA-Σ2 uses the fused rank-2 Triton for-

ward with a matched PyTorch backward. The phase-transition budget scales comparably to softmax at all scales tested, but without softmax's $\mathcal{O}(T^2)$ memory and compute.

H.5.2 KATA KERNEL MICROBENCHMARK

Forward and forward+backward wall-clock time on an NVIDIA H100 for the rank-two KATA-$\Sigma 2$ kernel, using a fused Triton forward and a PyTorch backward over saved chunk-boundary states in fp32:

Table 13: **KATA kernel microbenchmark.**

T	KATA-$\Sigma 2$ fwd	KATA-$\Sigma 2$ fwd+bwd	peak mem (bwd)
512	0.67 ms	3.58 ms	0.28 GB
1024	1.13 ms	6.55 ms	0.55 GB
2048	2.34 ms	12.07 ms	1.07 GB
4096	3.98 ms	23.49 ms	2.13 GB
8192	7.81 ms	46.66 ms	4.26 GB

The fused kernel scales linearly with T and stays near peak bandwidth; no per-token ψ tensor is written to HBM. The PyTorch backward rematerializes ψ once per chunk, avoiding an $\mathcal{O}(Tn_\psi)$ footprint. A fused backward kernel is left for future work.

H.5.3 DETAILED BENCHMARK NUMBERS

The tables below give the exact benchmark values. Tables 14 and 15 tabulate the forward and training-step curves in Figure 4; Table 16 reports the $B \times T$ forward-throughput grid; and Table 17 isolates the associative-scan and sequential-scan comparison at $B = 1$.

Table 14: **Forward latency versus sequence length.** Wall-clock milliseconds on one NVIDIA H100 (B=8, H=16, d_{head}=64, bf16); lower is better.

	Baselines		quadratic KATA-Mg $\mathcal{O}(T^2)$		linear-state KATA-Mg $\mathcal{O}(T)$	
T	FA-2	GDN	M1	M2	M1	M2
1K	0.09	0.54	0.07	0.07	2.68	0.68
2K	0.27	0.53	0.18	0.21	5.22	1.33
4K	0.93	0.66	0.58	0.72	10.34	2.70
8K	3.50	1.30	2.20	2.82	20.58	5.34
16K	13.78	2.59	8.92	10.62	41.14	10.68
32K	54.56	5.24	34.56	47.07	82.04	21.34
64K	234.2	10.67	150.8	168.7	164.1	42.70
128K	943.9	21.36	573.9	702.2	327.9	85.66

Table 15: **Training-step latency versus sequence length.** Forward-plus-backward milliseconds on one NVIDIA H100 (B=8, H=16, d_{head}=64, bf16); lower is better. Linear-state KATA-Mg uses the $\mathcal{O}(T)$ two-pass backward.

	Baselines		quadratic KATA-Mg $\mathcal{O}(T^2)$		linear-state KATA-Mg $\mathcal{O}(T)$	
T	FA-2	GDN	M1	M2	M1	M2
1K	0.44	2.62	0.69	0.72	10.14	3.37
2K	1.15	2.37	1.31	2.24	20.07	6.62
4K	3.70	2.35	4.01	7.51	39.93	13.11
8K	13.21	4.53	14.15	27.66	79.58	26.06
16K	49.97	9.07	53.66	106.4	159.1	52.14
32K	197.1	18.58	217.9	418.8	317.6	103.8

Table 16: **Forward throughput across the $B \times T$ grid.** Mtok/s on one NVIDIA H100 (H=16, d_{head}=64, bf16), sorted by $B{\cdot}T$. Tree-scan and linear-chunk use matched $d{\times}d$ states with C=128. Speedup is tree-scan over linear-chunk; higher is better.

B	T	$B{\cdot}T$	GDN	Linear-chunk	Tree-scan	speedup
1	2K	2K	2.5	4.9	**14.2**	2.9×
1	8K	8K	10.1	21.4	**44.2**	2.1×
4	2K	8K	9.1	19.6	**64.9**	3.3×
1	32K	32K	30.3	24.9	**49.2**	2.0×
4	8K	32K	38.1	37.6	**80.0**	2.1×
16	2K	32K	35.9	40.8	**93.3**	2.3×
1	64K	64K	30.6	25.4	**67.6**	2.7×
32	2K	64K	49.7	42.6	**99.0**	2.3×
4	32K	128K	45.9	39.2	**84.7**	2.2×
16	8K	128K	48.3	41.9	**100.5**	2.4×
4	64K	256K	45.3	39.6	**82.8**	2.1×
32	8K	256K	49.8	43.2	**103.1**	2.4×
16	32K	512K	47.6	41.7	**102.3**	2.5×
16	64K	1M	46.7	41.1	**88.3**	2.2×
32	32K	1M	48.6	42.3	**104.2**	2.5×
32	64K	2M	47.6	37.4	**89.4**	2.4×

Table 17: **Small-batch associative-scan throughput.** Forward Mtok/s at B=1 on one NVIDIA H100 (H=16, d_{head}=64, bf16). Tree-scan and linear-chunk use $\psi(\boldsymbol{x})$=$\boldsymbol{x}$, C=128, and matched $d{\times}d$ states; Gated DeltaNet is shown for reference. Speedup is tree-scan over linear-chunk; higher is better.

T	Gated DeltaNet (Mtok/s)	Linear-chunk (Mtok/s)	Tree-scan (Mtok/s)	speedup vs. chunk
2 048	2.6	6.7	**19.4**	2.9×
4 096	4.7	13.6	**39.9**	2.9×
8 192	8.1	23.4	**44.5**	1.9×
16 384	18.7	24.1	**47.6**	2.0×
32 768	30.4	24.9	**49.4**	2.0×
65 536	28.1	25.5	**67.5**	2.6×
131 072	30.8	25.8	**69.4**	2.7×

Table 18: **Training compute.** End-to-end throughput and accelerator time for each 15B-token 340M pretraining run; the final row aggregates all MQAR and overwrite experiments.

Run	GPU	k tok/s/GPU	GPU-hours
Softmax attention	H100	188	22.2
Gated DeltaNet	H100	144	28.9
KATA-M1 (Flash)	H100	190	21.9
KATA-M2 (Flash)	H100	190	21.9
DeltaKATA-M1	H100	77	54.1
MQAR + overwrite	A100	–	$\sim$4

H.6 TRAINING COMPUTE

Table 18 converts the measured end-to-end throughput into accelerator time using $15 \times 10^9/(r \times 3600)$ GPU-hours for throughput r in tokens/s/GPU. Under near-linear four-GPU scaling, the H100 totals correspond to approximately 5.5 hours for softmax attention, 7.2 hours for Gated DeltaNet, 5.5 hours for either Flash KATA model, and 13.5 hours for DeltaKATA-M1. The softmax run reaches approximately 760k tokens/s across four H100 GPUs.

Kernel-level attention gains are amortized across the full model, whose runtime also includes the MLPs, projections, RMSNorm, and optimizer work. Where applicable, we use FLA's fused recurrent kernels (Yang & Zhang, 2024), including the Gated DeltaNet and delta-rule recurrences, to reduce launch and intermediate-materialization overhead. The complete MQAR and overwrite benchmark suite is much smaller, requiring approximately 4 A100 GPU-hours in total.