LYNX: LINEARIZING TRANSFORMERS WITH TIMESCALE-SEPARATED MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Linearizing pretrained Transformers reduces long-context inference costs without pretraining from scratch, but preserving their capabilities under limited adaptation remains challenging. Our spectral analysis reveals low effective rank in the linearized baseline's recurrent state, with energy concentrated in a few directions. This limited use of the state's dimensions motivates a memory design with distinct retention timescales. We propose LYNX, a lightweight linearization method that combines sliding-window attention for exact recent access with timescale-separated recurrent states for historical information. The states jointly support association updates and fact retention, while a controller decouples retention from write allocation. A Landmark cache complements recurrent memory with exact access to selected historical tokens. Our analysis characterizes how retention shapes spectral concentration and gives sufficient conditions for the separated states to attain higher effective rank than a shared-retention baseline. Experiments on 8B-parameter LLaMA-family models evaluate general and long-context capabilities. LYNX achieves $71.79 \pm 0.33\%$ average accuracy across six reasoning and knowledge tasks, demonstrating general capability recovery. On long-context tasks, LYNX reaches a seven-task LongBench average of 26.07 and outperforms the compared linearization baselines across all eight Extended RULER task groups at 4K.

1 INTRODUCTION

Large language models (LLMs) can generate coherent text, answer questions, and solve reasoning tasks (Brown et al., 2020; Grattafiori et al., 2024). However, pretrained Transformer-based LLMs become increasingly costly to run as context length grows. Their softmax attention requires $O(n^2)$ computation for n tokens, while the key–value (KV) cache grows as $O(n)$ during decoding. Linear attention (Katharopoulos et al., 2020; Choromanski et al., 2020; Yang et al., 2023) reduces these costs by storing history in a fixed-size recurrent state, with $O(n)$ computation at fixed model width. Despite these efficiency gains, pretraining LLMs with linear attention from scratch remains costly and can involve trillions of training tokens (Van Nguyen et al., 2026). This cost has motivated methods that linearize pretrained Transformers through attention replacement and lightweight adaptation (Lan et al., 2025; Van Nguyen et al., 2026; Kuzina et al., 2026). These methods reuse pretrained weights to reduce training cost. However, recovering the source model's capabilities with limited adaptation remains a key challenge.

One aspect of this challenge is how effectively linearized models use recurrent memory. Sliding-window attention (SWA) (Zhang et al., 2025; Lan et al., 2025) supports recent access through a small softmax window, while out-of-window history is represented by the linear state S_t. Linear-attention models trained from scratch use large-scale pretraining to jointly learn their representations and recurrent state updates (Yang et al., 2023; 2025). Linearizing pretrained Transformers instead involves adapting representations learned with softmax attention to recurrent memory under a limited training budget. This raises a question: *how fully does the adapted model use the available state dimensions?* We investigate this through the effective rank of S_t, which measures how broadly its spectral energy is distributed. The linearized baseline's terminal state S_T has low effective rank, with most energy concentrated in a few directions. We hypothesize that limited adaptation produces states with lower effective rank than end-to-end linear-attention pretraining, leading to weaker long-context performance on tasks requiring access to distant information.

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.

To narrow this gap through effective-rank recovery, we propose LYNX, which linearizes pretrained Transformers through memory separation across timescales and exact access. SWA preserves exact recent access, while linear attention receives tokens after they leave the window. Within the linear path, a Medium state S_t^M supports dynamic updates through shorter retention timescales. A Long state S_t^L uses slower decay to retain stable facts. A lightweight controller separates retention from writing, and LoRA (Hu et al., 2021) adapts the pretrained representations. Queries read both states, with a Landmark cache providing additional exact access to selected historical tokens. In the historical diagnostic, our model's median effective rank at 4K is $2.08\times$ the linearized baseline's at the same state size (Figure 3).

Our main contributions are as follows:

- We identify low effective rank in the linearized baseline's recurrent state S_t, with most energy concentrated in a few directions. Our analysis shows how retention controls spectral concentration and effective rank, providing a theoretical motivation for timescale-separated states (Section 4).
- We introduce a lightweight Transformer linearization method that combines SWA and timescale-separated linear memory to jointly support recent access and distant fact retention. We characterize spectral conditions for higher effective rank than a shared-retention baseline (Section 5).
- Experiments demonstrate effective capability recovery on general reasoning and long-context tasks. LYNX improves average accuracy over Liger by an absolute 4.19% across six reasoning and knowledge tasks, and raises the seven-task LongBench average by 12.17–13.05 points over Liger and Lizard baselines (Section 6).

2 Related Work

2.1 Linear Attention and Recurrent Memory

Early linear-attention methods use kernel feature maps to compute attention in linear time without explicitly forming the attention matrix (Katharopoulos et al., 2020; Choromanski et al., 2020). Their recurrent form accumulates historical key–value associations in a fixed-size state S_t. RetNet (Sun et al., 2023) assigns fixed decay rates across heads, while GLA (Yang et al., 2023) uses input-dependent forgetting gates. In state-space models, Mamba (Gu & Dao, 2023) uses input-dependent selection, and Mamba-2 (Dao & Gu, 2024) connects structured state-space models with attention through state space duality. DeltaNet (Yang et al., 2024) uses delta-rule updates. Gated DeltaNet (Yang et al., 2025) adds adaptive forgetting, and Gated DeltaNet-2 (Hatamizadeh et al., 2026) further decouples erasure and writing with independent channel-wise gates. Kimi Linear (Team et al., 2025) introduces Kimi Delta Attention (KDA) with channel-wise forgetting and combines it with Multi-Head Latent Attention layers. However, costly pretraining motivates linearizing existing Transformers while preserving their capabilities through lightweight adaptation.

2.2 Post-training Transformer Linearization

Post-training linearization addresses this recovery problem through weight reuse, attention approximation, and adaptation. Hedgehog (Zhang et al., 2024) learns softmax-like feature maps, while SUPRA (Mercat et al., 2024) uptrains the replacement attention together with the original network. The Mamba in the Llama (Wang et al., 2024) initializes recurrent hybrids from pretrained projections and trains them through distillation. LoLCATs (Zhang et al., 2025) separates attention-output matching from LoRA adaptation. Lizard (Van Nguyen et al., 2026) also transfers attention outputs before language-modeling adaptation, using a hybrid of linear attention and Anchor Window Attention. In contrast, Liger (Lan et al., 2025) directly adapts a SWA–linear hybrid with LoRA. The Key to Going Linear (Kuzina et al., 2026) studies state-update choices with a frozen backbone. LoLA (McDermott et al., 2025) and STILL (Meng et al., 2026) cache selected historical tokens for exact access, while the remaining history is compressed into the recurrent state S_t. LYNX focuses on this state, using effective-rank analysis to guide memory organization across different retention timescales.

3 Problem Setup

3.1 Causal Softmax Attention

Given an input sequence $x_{1:N}$, the source projections produce $q_t, k_t \in \mathbb{R}^{d_k}$ and $v_t \in \mathbb{R}^{d_v}$. We write queries, keys, and values as column vectors and outputs as row vectors:

$$o_t^{\text{softmax}} = \frac{\sum_{i \le t} \exp(q_t^\top k_i / \sqrt{d_k}) v_i^\top}{\sum_{j \le t} \exp(q_t^\top k_j / \sqrt{d_k})}. \tag{1}$$

Each query accesses historical tokens through their individual key–value pairs. For N tokens at fixed model width, full-sequence attention requires $O(N^2)$ computation, while the KV cache grows as $O(N)$ during decoding. These costs limit efficient long-context inference.

3.2 Linear Attention and Recurrent State

Let $\tilde{q}_t = \varphi(q_t)$ and $\tilde{k}_t = \varphi(k_t)$, where d_φ is the feature dimension. Basic kernel linear attention (Katharopoulos et al., 2020; Choromanski et al., 2020) compresses historical associations into a fixed-size state:

$$S_t = S_{t-1} + \tilde{k}_t v_t^\top, \qquad z_t = z_{t-1} + \tilde{k}_t, \qquad o_t^{\text{linear}} = \frac{\tilde{q}_t^\top S_t}{\tilde{q}_t^\top z_t}. \tag{2}$$

The initial states are $S_0 = z_0 = 0$. The matrix $S_t \in \mathbb{R}^{d_\varphi \times d_v}$ stores key–value associations, while z_t provides the normalization state.

Gated DeltaNet (GDN) (Yang et al., 2025) combines decay with associative correction:

$$S_t = \alpha_t \big(I - \beta_t k_t k_t^\top\big) S_{t-1} + \beta_t k_t v_t^\top. \tag{3}$$

Here $S_t \in \mathbb{R}^{d_k \times d_v}$, α_t controls decay, and β_t controls the delta correction. The readout is $o_t = q_t^\top S_t$, with query/key preprocessing and fixed scaling implicit in the notation. Our implementation uses grouped key-channel decay inspired by KDA (Team et al., 2025). Its transition is $(I - \beta_t k_t k_t^\top) D_t$, where D_t is a diagonal decay matrix.

4 Understanding the Recurrent State in Linear Attention

4.1 What Does S_t Retain Beyond the Window?

Beyond the recent window, access to historical information relies on S_t. We first examine whether its rows represent distinct or overlapping directions. Large absolute cosine similarities indicate directional overlap; effective rank r_{eff} quantifies how broadly the state's spectral energy is distributed, accounting for both row directions and magnitudes.

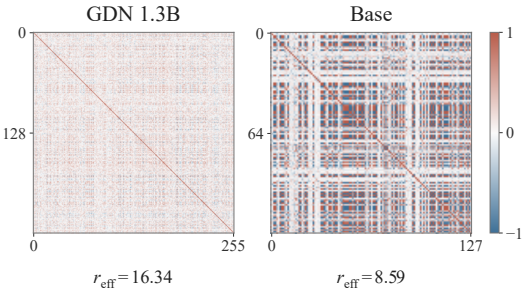


Figure 1: State-row cosine similarity. GDN is pretrained from scratch; Base is a linearized Transformer. Both process the same text prefix. Effective ranks are normalized.

Base denotes a pretrained Transformer linearized using GDN. Figure 1 shows stronger row correlations and lower normalized effective rank in the selected Base state than in the pretrained GDN reference. Its pronounced off-diagonal structure indicates overlapping row directions, while its lower effective rank (8.59 versus 16.34) reflects more concentrated spectral energy. For retrieval, the key question is whether this concentration makes historical associations harder to distinguish as new context arrives. This links our state analysis to the task of retaining and retrieving facts. We next examine how retention controls the distribution of state energy. Appendix F.1 gives the metric definitions, their connection to row similarity, and the sampling details.

4.2 Retention and Effective Rank

We isolate the effect of the decay coefficient α on effective rank using the GDN update in Equation 3 under balanced revisits. Let $U = [u_1, \ldots, u_r]$ and $W = [w_1, \ldots, w_r]$ have orthonormal columns, with $r \geq 2$. Starting from $S_0 = 0$, each of $m \geq 2$ rounds visits addresses $u_1, \ldots, u_r$ in the same order, writing values $c_\ell w_i$ in round ℓ, where $c_\ell > 0$ is shared across addresses. The coefficients $\alpha \in (0, 1]$ and $\beta \in (0, 1)$ are constant. Each revisit partially overwrites the previous association. At $T = mr$ memory-write steps, let $d_i = r - i$. The state and its spectral energies follow from

$$C_m(\alpha, \beta) = \beta \sum_{\ell=1}^{m} c_\ell \left[(1-\beta)\alpha^r\right]^{m-\ell}, \qquad p_i(\alpha) = \frac{\alpha^{2d_i}}{\sum_j \alpha^{2d_j}}, \tag{4}$$

$$S_T(\alpha) = C_m(\alpha, \beta) U \operatorname{diag}(\alpha^{d_1}, \ldots, \alpha^{d_r}) W^\top. \tag{5}$$

Appendix E.1 derives this expression with the full overwrite term.

Lemma 1. *(Proof in Appendix E) Under the balanced-revisit model above, let $\theta = -2 \log \alpha$. For $0 < \alpha < 1$,*

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \log r_{\text{eff}}(S_T(\alpha)) = \frac{2\theta}{\alpha} \operatorname{Var}_{p(\alpha)}(d) > 0. \tag{6}$$

Hence, with matched inputs and β, increasing retention from α_B to α_A ($0 < \alpha_B < \alpha_A \leq 1$) gives $r_{\text{eff}}(S_T(\alpha_A)) > r_{\text{eff}}(S_T(\alpha_B))$.

Repeated writing and overwriting contribute the common amplitude C_m, while longer retention balances spectral energy and increases effective rank under these assumptions. These findings motivate separating retention timescales to jointly support association updates and stable-fact retention.

5 Linearization with Timescale-Separated Memory

LYNX combines exact recent access with timescale-separated recurrent states to support historical updates and fact retention. It reuses pretrained Transformer projections and introduces a lightweight controller for state-specific writing and retention.

5.1 Memory Design and Operation

Motivated by the spectral concentration observed in the linearized state, we introduce Medium and Long states with different retention timescales. Both receive the same historical stream and jointly support association updates and fact retention. SWA preserves exact recent access; non-sink tokens enter recurrent memory only after leaving its window.

Both states preserve the full query/key address space and split the value dimension equally, allowing a shared query to produce a full-width output. At input time t, the states for each recurrent head are

$$S_t^M, S_t^L \in \mathbb{R}^{d_k \times (d_v/2)}, \qquad \bar{S}_t = [S_t^M \mid S_t^L]. \tag{7}$$

Medium and Long differ in their retention ranges, not in which tokens they receive. Adaptive write allocation controls how strongly each token updates the states, while their distinct retention timescales govern persistence. Figure 2 summarizes this organization and its joint readout.

Controller and write routing. The controller determines *how long information persists*, *how strongly it is written*, and *how writing is allocated*. Separate heads predict these controls from shared low-dimensional features of the token leaving the window at time t:

$$\alpha_t^b = 2^{-1/\tau_t^b}, \qquad \beta_t^b = s\beta_t p_t^b, \qquad p_t^M + p_t^L = 1, \tag{8}$$

where $b \in \{M, L\}$ and s scales writing. Head and retention-group indices are omitted for clarity. Retention uses input-dependent nominal half-lives within non-overlapping ranges: Medium's upper bound is below Long's lower bound (Appendix B). The total strength β_t is shared across states; router weights p_t^b divide it between them.

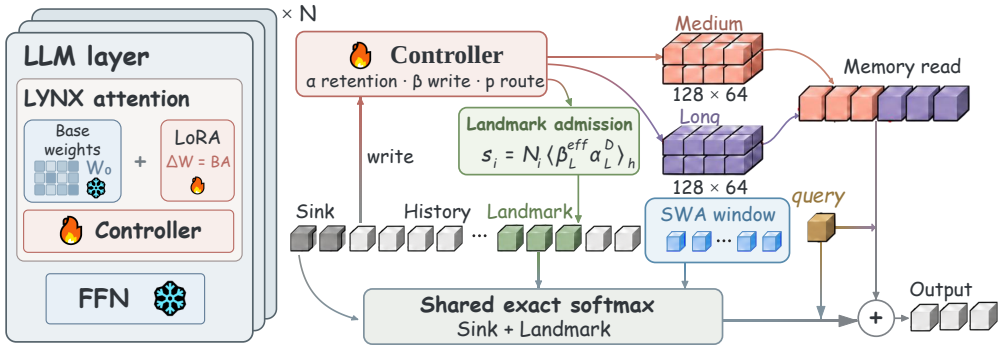


Figure 2: Overview of LYNX. Left: pretrained weights remain frozen while LoRA and the controller are trained. Right: the controller governs retention and write allocation in Medium and Long states; Landmark selection preserves selected historical K/V pairs. Sink, recent-window, and Landmark tokens share one exact softmax. The query reads both recurrent states, whose outputs are concatenated and added to the exact-attention output.

Memory writing. The non-sink token at position $i(t)$, which leaves the window at time t, updates both states:

$$S_t^b = \mathcal{U}_b\left(S_{t-1}^b, k_{i(t)}, v_{i(t)}^b; \alpha_t^b, \beta_t^b\right). \tag{9}$$

If no token exits, $S_t^b = S_{t-1}^b$, with no decay or writing. The states share $k_{i(t)}$ and receive separate value components $v_{i(t)}^b$. The update $\mathcal{U}_b$ uses the grouped-decay delta rule described in Section 3.2, with state-specific retention and write strength.

Landmark selection. Landmark preserves exact K/V pairs for selected historical tokens. Selection combines Long-state *write-and-retention priority* with *address novelty*, balancing persistence against redundant cache entries. For the exiting token $i = i(t)$, define

$$P_i = \mathrm{mean}_h\left[\beta_{t,h}^L (\bar{\alpha}_{t,h}^L)^D\right], \qquad N_i = \mathrm{clip}\left(1 - \max_{j \in \mathcal{L}_{t-1}} z_i^\top z_j, \epsilon, 1\right). \tag{10}$$

P_i favors tokens with stronger Long-state writes and slower decay, while N_i downweights candidates whose key directions are already represented in the cache. Here $\bar{\alpha}_{t,h}^L = \exp(\frac{1}{G}\sum_g \log \alpha_{t,h,g}^L)$ averages retention geometrically across groups, D is a fixed horizon, and z_i is a normalized key summary. We set $N_i = 1$ for an empty cache; $\epsilon > 0$ is the novelty floor. The cache retains the top C_{lm} scores:

$$\sigma_i = P_i N_i, \qquad \mathcal{L}_t = \mathrm{TopK}_{C_{\mathrm{lm}}}\left(\mathcal{L}_{t-1} \cup \{i(t)\}; \sigma\right). \tag{11}$$

The cache stays unchanged without an exiting token; scores are frozen at exit. Selected tokens also update both recurrent states, so exact storage complements their integration into compressed history. Appendix A.1 specifies key summaries, scoring constants, and implementation details.

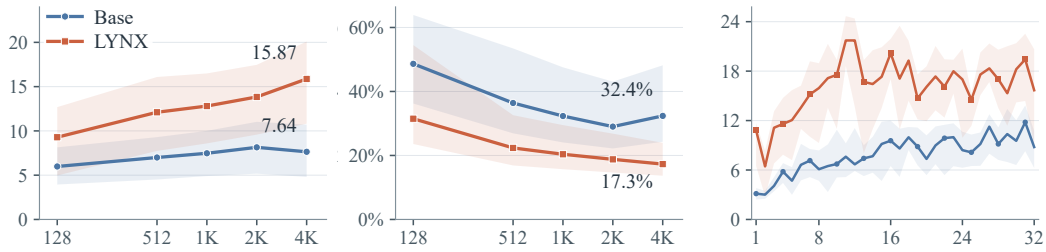
Query readout. At input time t, the exact path reads sink and recent tokens together with any selected Landmarks, indexed by $\mathcal{A}_t$, through one causal softmax:

$$o_t^{\mathrm{exact}} = \frac{\sum_{j \in \mathcal{A}_t} \exp(\hat{q}_t^\top \hat{k}_j / \sqrt{d_k}) v_j^\top}{\sum_{j \in \mathcal{A}_t} \exp(\hat{q}_t^\top \hat{k}_j / \sqrt{d_k})}. \tag{12}$$

Here $\hat{q}_t, \hat{k}_j$ are RoPE-transformed queries and keys (Su et al., 2024). The query also reads both recurrent states. Their outputs are concatenated along the value dimension and added to the exact-access output:

$$o_t^{\mathrm{mem}} = \mathrm{concat}\left(q_t^\top S_t^M, q_t^\top S_t^L\right), \qquad o_t = o_t^{\mathrm{exact}} + o_t^{\mathrm{mem}}. \tag{13}$$

This joint readout combines historical information retained at different timescales with recent context. The router controls *write allocation*; both states contribute directly to readout without an additional routing gate.



(a) Effective rank (b) Leading-mode energy (c) Layerwise effective rank

Figure 3: Effective rank (a,c) and leading-mode energy fraction (b) versus context length (a–b, log scale) or layer at 4K (c). LYNX has higher effective rank and lower leading-mode energy than Base, indicating less concentrated state spectra. Its effective rank increases across sampled context lengths, while Base declines beyond 2K. The advantage extends across all 32 layers at 4K.

5.2 Effective Rank and Memory Utilization

We examine whether timescale-separated states distribute spectral energy across a broader range of directions.

Figure 3 compares states of equal dimensions. LYNX has higher effective rank r_{eff} and a lower leading-mode energy fraction p_1 across sampled lengths; its rank advantage spans all 32 layers at 4K. Figure 4 provides a complementary view: LYNX exhibits weaker off-diagonal row correlations than Base on the same input. Together, these observations show less overlap among state-row directions and a broader spectral energy distribution. These improvements reflect broader use of the available state dimensions at a fixed matrix size.

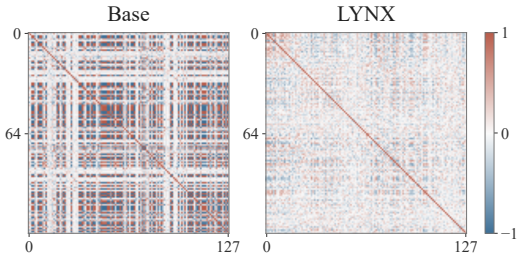


Figure 4: State-row cosine similarity after the same input. Our state concatenates Medium and Long along the value dimension.

To interpret this spectral broadening, we give a sufficient condition for higher effective rank. Let $S_T^A = S_T$ be the LYNX state and S_T^B the Base state on the same input, with equal matrix dimensions and the same number $r \geq 2$ of positive singular values. Order each state's squared singular values in descending order and define

$$e_i^X = \sigma_i(S_T^X)^2, \qquad p_i^X = \frac{e_i^X}{\sum_j e_j^X}, \qquad g_i = \frac{e_i^A}{e_i^B}, \quad X \in \{A, B\}. \tag{14}$$

Here g_i compares the same ordered spectral position in the two states. Larger ratios at lower-energy positions indicate a relative shift away from the dominant components.

Lemma 2. *(Proof in Appendix E) For the ordered spectra above, suppose the relative energy ratios satisfy*

$$0 < g_1 \leq g_2 \leq \cdots \leq g_r. \tag{15}$$

Then $\sum_{i=1}^k p_i^A \leq \sum_{i=1}^k p_i^B$ *for* $1 \leq k < r$, *and*

$$r_{\text{eff}}(S_T^A) \geq r_{\text{eff}}(S_T^B), \qquad p_1^A \leq p_1^B. \tag{16}$$

Both inequalities in Equation 16 are strict when the ratios are not all equal.

Lemma 2 links relative gains in lower-energy spectral components to higher effective rank. Appendix E.2 gives a retention-controlled case with repeated writes and overwrites, where a strict rank advantage survives sufficiently small perturbations. Separate Medium/Long retention controls thus support broader use of the available state dimensions under this spectral condition.

6 Overall Linearization Results

6.1 Experimental Setup

Models and evaluation. We linearize LLaMA-3-8B and LLaMA-3.1-8B (Grattafiori et al., 2024). LLaMA-3-8B is evaluated on six reasoning and knowledge tasks, while LLaMA-3.1-8B is evaluated on LongBench (Bai et al., 2024) and RULER (Hsieh et al., 2024). Linearization uses Alpaca instruction data (Taori et al., 2023). The six tasks are PIQA (Bisk et al., 2020), ARC-e/c (Clark et al., 2018), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), and MMLU (Hendrycks et al., 2020), evaluated using `lm-eval` (Gao et al., 2023)[1] with zero-shot prompts for the first five and five-shot prompts for MMLU. We report their equal-weight average as mean $\pm$ sample standard deviation across three training seeds. Long-context evaluation covers seven LongBench tasks, S-NIAH at multiple context lengths, and eight Extended RULER tasks.

Adaptation for long-context tasks. For long-context tasks, we additionally adapt the model with supervision on the complete answer and its end-of-sequence (EOS) token, masking context and question labels. This objective trains answer generation and termination from the supplied context. We then train a causal K/V retaining head to calibrate Landmark priorities while freezing the adapted model. Appendices B and C detail training, supervision, checkpoints, and long-context protocols; Appendices D.1–D.2 report per-seed general-task and ablation results.

6.2 Overall Results

Table 1: Accuracy (%) on six reasoning and knowledge benchmarks. We compare LLaMA-3-8B linearized models with the source Transformer and independently pretrained efficient models. Avg. denotes the six-task mean; bold scores mark the highest mean in each column within the Linearized group.

Model	Tok. (B)	PIQA	ARC-e	ARC-c	Hella.	Wino.	MMLU	Avg.
Transformer								
LLaMA-3-8B	—	79.90	80.10	53.30	79.10	73.10	66.60	72.02
Native Efficient Models								
Mamba-7B†	1200	81	77.50	46.70	77.90	71.80	33.30	64.70
Hawk-7B†	300	80	74.40	45.90	77.60	69.90	35	63.80
Griffin-7B†	300	81	75.40	47.90	78.60	72.60	39.30	65.80
Linearized								
Mamba2-LLaMA	20	76.80	74.10	48	70.80	58.60	43.20	61.92
LoLCATs	0.04	**80.90**	**81.70**	54.90	79.70	74.10	52.80	70.68
Liger-GLA	0.02	80.30	81.10	52.50	76.30	72	43.40	67.60
LYNX	0.01	$80.27_{\pm 0.60}$	$81.50_{\pm 0.27}$	$\mathbf{55.26}_{\pm 0.89}$	$\mathbf{80.30}_{\pm 0.12}$	$\mathbf{75.14}_{\pm 0.44}$	$\mathbf{58.29}_{\pm 0.54}$	$\mathbf{71.79}_{\pm 0.33}$

Tok. (B): billions of pretraining tokens for † models, adaptation tokens otherwise. LYNX: mean $\pm$ sample SD across three training seeds.

Table 1 compares LLaMA-3-8B linearizations with independently pretrained 7B references, including Hawk and Griffin (De et al., 2024). Public scores and token budgets follow Van Nguyen et al. (2026), Table 1; independent models report pretraining tokens and linearized models report adaptation tokens. Our row reports multi-seed statistics for 1,000-step LLaMA-3-8B checkpoints. LoLCATs and Liger use 64-token windows (Zhang et al., 2025; Lan et al., 2025). Training and evaluation settings retain their source protocols.

General-task results. Our six-task mean is $71.79 \pm 0.33\%$ across three training seeds, above Liger-GLA's 67.60 and LoLCATs' 70.68, versus 72.02 for the full-attention reference. Relative to Liger, gains on MMLU (43.40 to 58.29) and ARC-c (52.50 to 55.26) support recovery of knowledge-intensive reasoning. Higher HellaSwag and WinoGrande scores extend this pattern to contextual interpretation and candidate discrimination, while PIQA remains similar. Together, these results support the combined exact-attention and timescale-separated design for general capability recovery.

[1] `https://github.com/EleutherAI/lm-evaluation-harness`

LongBench results. Figure 5 uses LoLA's seven-task selection and reported LoLCATs scores (McDermott et al., 2025). These tasks assess whether a model can use contextual evidence to answer questions and identify relevant information.

Our seven-task mean is 26.07, versus 19.40 for GDN and 13.02 for Lizard. We outperform GDN on six tasks, including HotpotQA, 2WikiMQA, and MuSiQue. These multi-hop tasks require combining evidence from different parts of the context, connecting the gains to the historical-information access targeted by our design. Exact attention preserves access to selected tokens, while Medium and Long states integrate historical information at different timescales. Their joint readout supports the use of this evidence in answering questions. GDN remains stronger on TriviaQA; the clearest gains are on tasks requiring contextual evidence integration. This pattern supports retaining both selected token-level evidence and recurrent summaries of the wider context for long-context understanding.

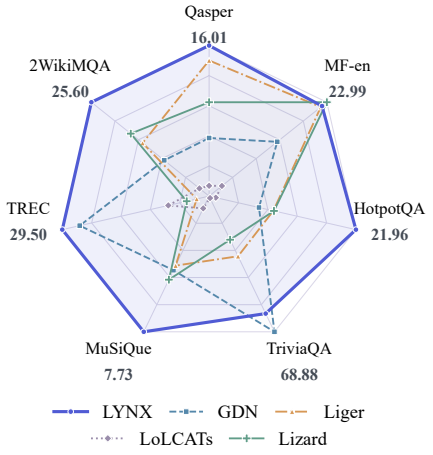


Figure 5: LongBench scores scaled from zero to the labeled per-task maxima.

6.3 S-NIAH Retrieval

Table 2 evaluates S-NIAH-1/2/3 from RULER (Hsieh et al., 2024) at 0.5K/1K/2K/4K context lengths; LoLA scores follow McDermott et al. (2025), Table 1. These tasks test whether information remains retrievable as the surrounding context grows.

LYNX (64, 64) maintains 100% on S-NIAH-1 across all lengths and reaches 100.0/91.4/5.0 on S-NIAH-1/2/3 at 4K, compared with 99.6/48.4/2.2 for LYNX (64, 16). On S-NIAH-2, Lizard declines from 81.6 to 39.6 as context length increases. The strong S-NIAH-1/2 results support our design's ability to keep historical information accessible beyond the recent window, combining recurrent states with exact access to selected historical tokens. S-NIAH-3 remains difficult for all compared models, showing that precise retrieval is still sensitive to the type of information being retained.

Table 2: S-NIAH recall accuracy (%) across context lengths. Parentheses denote cache sizes in tokens: (local, historical) for LYNX, LoLA, and Lizard, and (local) for the other methods. Dashes denote unavailable scores.

Model	S-NIAH-1				S-NIAH-2				S-NIAH-3			
	0.5K	1K	2K	4K	0.5K	1K	2K	4K	0.5K	1K	2K	4K
Liger (128)	25.2	9.0	0.6	0.0	24.6	13.8	6.6	2.0	26.2	11.4	6.4	2.4
GDN (128)	25.8	9.2	2.6	1.4	22.4	10.6	4.4	2.2	20.8	12.6	6.4	3.4
Lizard (64, 64)	98.6	99.2	99.6	100.0	81.6	67.0	53.4	39.6	0.0	0.0	0.0	0.0
LoLA (64, 64)	99.0	95.4	79.4	69.4	100.0	39.4	3.0	—	99.8	7.4	1.6	—
LYNX (64, 16)	100.0	99.8	100.0	99.6	100.0	97.2	49.6	48.4	81.8	3.0	1.0	2.2
LYNX (64, 64)	100.0	100.0	100.0	100.0	100.0	100.0	99.4	91.4	100.0	75.8	27.2	5.0

6.4 Extended RULER

Extended RULER evaluates retrieval, tracking, word extraction, and question answering (Hsieh et al., 2024; McDermott et al., 2025). Figure 6 compares our model with Liger and Lizard across eight task groups at 4K.

Our model leads on all eight task groups, reaching 63.76 on variable tracking, 46.55 on multi-value retrieval, and 43.30 on multi-query retrieval. These tasks require tracking changing associations and retaining multiple items, supporting our joint design for updating and accessing historical information. Multi-key retrieval (7.93) and common-word extraction (8.74) remain challenging.

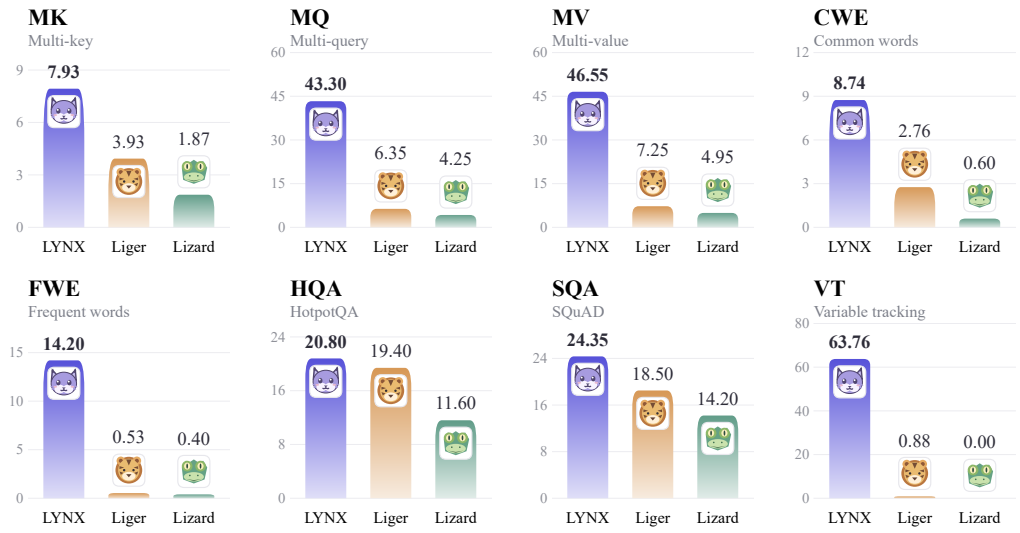


Figure 6: Extended RULER at 4K (scores in %); each panel uses its own zero-based scale. MK averages three multi-key tasks. All three models use 128 exact-cache tokens; Liger and Lizard include our retaining-head adaptation.

6.5 Ablation Study

Figure 7 presents an MMLU ablation of five configurations, all with SWA. S (Single) uses one recurrent state. The two-state variants use shared retention in T (Tied) and independent dynamic controls within a common timescale range in F (Free). T+LM8 combines tied retention with eight Landmarks, while LYNX uses timescale-separated states and 16 Landmarks. These configurations vary in both controls and cache capacity.

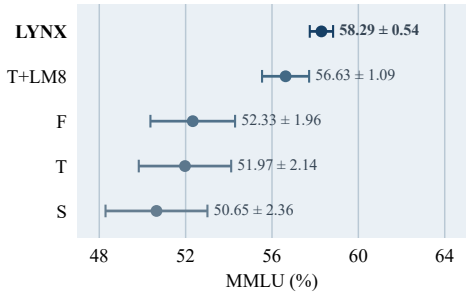


Figure 7: Five-shot MMLU after 1,000 steps (mean ± SD, three seeds). Whiskers indicate one standard deviation.

The two-state variants raise mean MMLU accuracy from 50.65% (S) to 51.97% (T) and 52.33% (F), with differences small relative to seed variation. Gains are larger with Landmarks: T+LM8 reaches 56.63%, while LYNX achieves $58.29 \pm 0.54\%$, the highest mean and lowest variation. The full model's performance supports the effectiveness of jointly using timescale-separated states and Landmark selection to recover knowledge-intensive task performance after linearization.

7 Conclusion

We study Transformer linearization through the effective rank of recurrent states and identify concentrated state spectra in the linearized baseline. This motivates a design that combines exact recent access, timescale-separated Medium and Long states, and Landmark selection for historical tokens. Decoupled writing and retention let the two states jointly accommodate changing associations and persistent facts. Spectral analysis shows broader state utilization, while experiments on LLaMA-3-8B and LLaMA-3.1-8B demonstrate strong general-task recovery and improved long-context performance. Together, these findings highlight the value of organizing bounded memory around complementary update, retention, and access mechanisms for effective Transformer linearization.

AI USE STATEMENT

We used generative AI tools to assist with language editing and polishing, literature retrieval and discovery, and the development of mathematical proofs. The proof-related assistance included refining assumptions, exploring derivations, and improving the presentation of arguments. The authors take responsibility for the final content, including the accuracy of citations and the correctness of mathematical claims and proofs.

REPRODUCIBILITY STATEMENT

Appendix A describes the state updates, Landmark selection, exact readout, and causal cache timing. Appendix B reports training settings, supervision objectives, and evaluation procedures, while Appendix C documents long-context results and their evaluation configurations. Appendix D provides per-seed results and routing diagnostics. The assumptions and complete proofs of both lemmas appear in Appendix E. Appendix F specifies the spectral metrics, normalization, checkpoint provenance, and sampling procedures used in the diagnostic figures.

REFERENCES

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Koenraad MR Audenaert. A sharp continuity estimate for the von neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, 2007.

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

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

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Soham De, Samuel L Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

Leo Gao, Jonathan Tow, Baber Abbasi, et al. A framework for few-shot language model evaluation, 2023. URL https://zenodo.org/records/10256836.

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

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yuxiang Huang, Binhang Yuan, Xu Han, Chaojun Xiao, and Zhiyuan Liu. Locret: Enhancing eviction in long-context llm inference with trained retaining heads on consumer-grade devices. *arXiv preprint arXiv:2410.01805*, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Anna Kuzina, Paul N Whatmough, and Babak Ehteshami Bejnordi. The key to going linear: Analysis-driven transformer linearization. *arXiv preprint arXiv:2607.07706*, 2026.

Disen Lan, Weigao Sun, Jiaxi Hu, Jusen Du, and Yu Cheng. Liger: Linearizing large language models to gated recurrent structures. *arXiv preprint arXiv:2503.01496*, 2025.

Yuan Li and Paul Busch. Von neumann entropy and majorization. *Journal of Mathematical Analysis and Applications*, 408(1):384–393, 2013.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Awadallah, and Dragomir Radev. Dyle: Dynamic latent extraction for abstractive long-input summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1687–1698, 2022.

Luke McDermott, Robert W Heath Jr, and Rahul Parhi. Lola: Low-rank linear attention with sparse caching. *arXiv preprint arXiv:2505.23666*, 2025.

Weikang Meng, Liangyu Huo, Yadan Luo, Jiawen Guan, Jingyi Zhang, Yingjian Li, and Zheng Zhang. Still: Selecting tokens for intra-layer hybrid attention to linearize llms. *arXiv preprint arXiv:2602.02180*, 2026.

Jean Mercat, Igor Vasiljevic, Sedrick Keh, Kushal Arora, Achal Dave, Adrien Gaidon, and Thomas Kollar. Linearizing large language models. *arXiv preprint arXiv:2405.06640*, 2024.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

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

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.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, et al. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

Chien Van Nguyen, Huy Huu Nguyen, Ruiyi Zhang, Hanieh Deilamsalehy, Puneet Mathur, Viet Dac Lai, Haoliang Wang, Jayakumar Subramanian, Ryan A Rossi, Trung Bui, et al. Lizard: An efficient linearization framework for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34935–34948, 2026.

Junxiong Wang, Daniele Paliotta, Avner May, Alexander M Rush, and Tri Dao. The mamba in the llama: Distilling and accelerating hybrid models. *Advances in Neural Information Processing Systems*, 37:62432–62457, 2024.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *International Conference on Learning Representations*, volume 2025, pp. 29687–29707, 2025.

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

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. In *International Conference on Learning Representations*, volume 2024, pp. 53551–53580, 2024.

Michael Zhang, Simran Arora, Rahul Chalamala, Benjamin Spector, Alan Wu, Krithik Ramesh, Aaryan Singhal, and Christopher Ré. Lolcats: On low-rank linearizing of large language models. In *International Conference on Learning Representations*, volume 2025, pp. 46200–46253, 2025.

A METHOD AND IMPLEMENTATION DETAILS

Each recurrent head maintains two 128×64 states with a shared 128-dimensional query/key address space. Their outputs are concatenated and added to the exact-attention output. With head index h and retention-group index g, the full controller notation is $\alpha^b_{t,h,g} = 2^{-1/\tau^b_{t,h,g}}$ and $\beta^b_{t,h} = s\beta_{t,h}p^b_{t,h}$, with $p^M_{t,h} + p^L_{t,h} = 1$ for $b \in \{M, L\}$. These controls are computed from token $i(t)$ leaving the window at input time t. At each input step:

1. Project the current input into queries, keys, and values. Apply each branch's positional transformations.
2. Admit a non-sink token when it exits the recent window. At input time t, apply $\mathcal{U}_M$ and $\mathcal{U}_L$ to its two value subspaces. If no token exits, retain both states and the Landmark cache unchanged.
3. If Landmark is enabled, compute the candidate's Long priority and address novelty. Freeze its score and update the bounded historical set.
4. Read both states with the current query and concatenate their value outputs. Add the exact softmax output over sink, recent, and Landmark tokens.

Sink tokens never enter recurrent memory. Chunk prefill and token-wise decoding follow the same window-exit events and eligibility boundary.

A.1 LANDMARK SCORING AND EXACT READOUT

Each layer selects up to C_{lm} historical K/V pairs using Long-state write priority and address novelty. Let $i = i(t)$ denote the original position of the token leaving the recent window, h index recurrent heads, and g index retention groups. Averaging log-retention coefficients across groups gives the Long priority

$$\bar{\ell}^L_{t,h} = \frac{1}{G}\sum_{g=1}^{G} \log \alpha^L_{t,h,g}, \qquad P_i = \frac{1}{H}\sum_{h=1}^{H} \beta^L_{t,h} \exp\!\left(D\bar{\ell}^L_{t,h}\right), \qquad D = 2048. \tag{17}$$

Here $H = 32$, $G = 8$, and $\beta^L_{t,h} = s\beta_{t,h}p^L_{t,h}$ is the effective Long write strength. The horizon D weights writing by its retained fraction.

We average the memory-branch key over $H_{\mathrm{KV}} = 8$ KV heads and normalize it to obtain an address summary:

$$\bar{k}^{\mathrm{mem}}_i = \frac{1}{H_{\mathrm{KV}}}\sum_{a=1}^{H_{\mathrm{KV}}} k^{\mathrm{mem}}_{i,a}, \qquad z_i = \frac{\bar{k}^{\mathrm{mem}}_i}{\max(\|\bar{k}^{\mathrm{mem}}_i\|_2, 10^{-6})}. \tag{18}$$

The address summary z_i defines novelty as

$$N_i = \begin{cases} 1, & \mathcal{L}_{t-1} = \varnothing, \\ \mathrm{clip}\left(1 - \max_{j \in \mathcal{L}_{t-1}} z_i^\top z_j,\ 0.05,\ 1\right), & \text{otherwise.} \end{cases} \tag{19}$$

Equation 11 retains the highest-scoring entries among the cache and newly eligible candidate. Scores are fixed at window exit, with ties resolved in favor of earlier tokens. Selection uses detached FP32 tensors; selected K/V pairs remain differentiable. Memory keys use RoPE at their original positions when enabled.

Landmark, sink, and recent tokens form the exact-attention set

$$\mathcal{A}_t = \mathcal{I}^{\mathrm{sink}}_t \cup \mathcal{I}^{\mathrm{recent}}_t \cup \mathcal{L}_t. \tag{20}$$

The augmented set uses the single softmax in Equation 12. Its output is added to the memory readout in Equation 13. Selected tokens also update the linear state. The exact KV budget is $8 + 56 + C_{\mathrm{lm}}$.

A.2 Supervised Retaining Head

Long-context adaptation adds a causal retaining head, following the frozen-model selector approach of Locret (Huang et al., 2024) and direct evidence supervision in DYLE (Mao et al., 2022). Each layer has a $256 \to 16 \to 1$ MLP with SiLU activation and biases, shared across its eight KV heads. For candidate i, it separately ℓ_2-normalizes each head's key and value (floor 10^{-6}), concatenates them, and scales by $\sqrt{128}$. The mean MLP output across KV heads is a scalar logit a_i. The 32 layer-specific MLPs contain 132,128 parameters in total.

The head corrects the Long priority from Equation 17:

$$g_i = 6\tanh(a_i/6), \qquad \widetilde{P}_i = P_i e^{g_i} + 10^{-6}\max(e^{g_i} - 1, 0), \qquad \widetilde{\sigma}_i = \widetilde{P}_i N_i. \tag{21}$$

The correction preserves novelty, causal eligibility, cache capacity, and recurrent writes. Zero-initializing the output layer recovers $P_i N_i$.

Training labels all tokens in annotated evidence records as positive; eligible background tokens are negative. Sink and unexpired recent tokens are ignored. At each layer, the loss averages positive-token binary cross-entropy and binary cross-entropy on the 64 highest-logit negatives, with equal weight; examples without positives use only the negative term. Layer losses are averaged. Positives contribute even when not admitted to the cache. K/V features are detached and only the retaining heads receive gradients. Inference uses causal K/V features without gold evidence or the later question.

B Experimental Settings

Table 3 summarizes the general-task training and the two additional long-context adaptation stages. General-task evaluation uses 1,000-step checkpoints, each trained on 8,192,000 input tokens. The main configuration has eight retention groups, memory RoPE, and a rank-16 controller. Medium half-lives lie in $[64, 1024]$ (initially 512); Long half-lives lie in $[2048, 16384]$ (initially 8192).

Table 3: Training settings. The last two columns describe the additional stages used by the LYNX long-context checkpoints.

Setting	General tasks	Answer + EOS	Retaining head
Backbone	LLaMA-3-8B	LLaMA-3.1-8B	Frozen preceding model
Evaluated stage	Step 1,000	200 steps	200 steps
Optimizer	AdamW	AdamW	AdamW
Learning rate	5×10^{-4}	10^{-4}	10^{-3}
Weight decay	0	0.01	0.01
Schedule	Cosine	Warmup–cosine	Warmup–cosine
Warmup steps	200	20	20
Global batch size	8	1	1
Context length	1,024	2K then 4K	Matching context prefixes
LoRA rank / scale	8 / 16	8 / 16	Frozen
Trainable parameters	9,207,808	9,207,808	132,128

General-task training uses Alpaca-cleaned answer supervision, gradient clipping at 1.0, and a 5,000-step cosine horizon. Its micro-batch size is four with two accumulation steps. Additional adaptation also clips gradients at 1.0 and uses a 200-step horizon for each stage.

B.1 Long-Context Supervision

Answer-plus-EOS training uses 200 examples: 20% natural-language, 20% numeric retrieval with repeated background, 20% numeric retrieval in articles, and 40% UUID retrieval in articles. The first 20% of steps use a 2K context limit and the remainder use 4K. Documents and key/value pools are separated across training, development, and confirmation splits. For retrieval examples, context and question labels are masked. With answer tokens $y_{1:K}$ and context/question x, the loss is

$$\mathcal{L}_{\text{ret}} = -\frac{1}{K}\sum_{j=1}^{K} \log p_\theta(y_j \mid x, y_{<j}) - \log p_\theta(\text{EOS} \mid x, y_{1:K}). \tag{22}$$

Natural-language examples use ordinary next-token cross-entropy. The retaining-head stage reuses the context prefixes with questions and answers removed, freezes the adapted model, and optimizes the evidence selection loss in Appendix A.2.

B.2 Evaluation and Budget Accounting

The six general tasks use `lm-eval` 0.4.9.2 with a maximum context of 4,096 tokens. PIQA, ARC-e, ARC-c, HellaSwag, and WinoGrande use zero-shot evaluation; MMLU uses five-shot evaluation. ARC-c and HellaSwag use normalized accuracy, and the remaining tasks use accuracy. All 30,737 questions are included. We average unrounded task scores before displaying two decimal places: the three-seed mean and sample standard deviation are $71.79 \pm 0.33\%$.

For memory accounting, B_S counts recurrent-matrix elements and B_{KV} counts exact KV entries. These budgets are separate.

C Complete Long-Context Results

Table 4 distinguishes the checkpoint families. All LYNX rows use SWA $8 + 56$, with separate checkpoints for the two historical-cache capacities.

Table 4: LYNX evaluation configurations. A200 and H200 denote 200-step answer-plus-EOS and retaining-head adaptation. Exact KV includes SWA.

Evaluation	Backbone	Parent step	Adaptation	Landmark / exact KV
Six general tasks	LLaMA-3-8B	1,000	None	16 / 80
S-NIAH (64,16)	LLaMA-3.1-8B	5,000	A200 + H200	16 / 80
LongBench / RULER	LLaMA-3.1-8B	5,000	A200 + H200	64 / 128

S-NIAH cache notation and scoring. In Table 2, pairs denote local and historical cache capacities, while (128) denotes total cache tokens. LoLA's local cache is its sliding window; ours contains eight sink and 56 recent tokens. RH denotes our retaining-head adaptation after answer-plus-EOS training. Our scores measure reference-substring accuracy over 500 samples per task and context length.

Local S-NIAH baselines at KV128. We reevaluate existing LLaMA-3.1-8B Base checkpoints on S-NIAH-1/2/3 at 512, 1,024, 2,048, and 4,096 tokens, with 500 examples per cell. Liger uses a 128-token sliding window at inference, expanded from its 64-token training window. Our local Paper-GDN reproduction uses the DCLM-edu (Allal et al., 2025) 2,441-step checkpoint, expanding its cache from 8 sink plus 56 recent tokens to 8 sink plus 120 recent tokens. Both evaluations use greedy decoding and case-insensitive reference-substring accuracy, with no additional training. GDN denotes our reproduction of Kuzina et al. (2026).

Local Lizard retaining adaptation. We extend the locally reproduced Lizard model on LLaMA-3.1-8B Base with 8 sink, 56 recent, and 64 historical cache tokens. Its recurrent numerator and normalizer states and four denominator-only meta logits are retained. Answer-plus-EOS adaptation trains 4,718,592 LoRA parameters for 200 steps at learning rate 10^{-4}. A further 200 steps train a 132,128-parameter retaining head with causal K/V evidence supervision at learning rate 10^{-3}, with the parent frozen. Both stages use 20 warmup steps; together they process 1,431,022 input tokens. LongBench uses 1,350 examples across seven tasks; RULER uses 500 examples per task. The S-NIAH sweep evaluates the same adapted model at 512/1,024/2,048/4,096 tokens, with 500 examples per task and length. Each length uses an independently fixed sample set, case-insensitive reference-substring scoring, a 128-token generation limit, and no input truncation.

Local seven-task comparison. The GDN row in Table 5 is our local Paper-GDN reproduction, adapted on DCLM-edu for 2,441 steps (9,998,336 input tokens), with 64 exact-cache tokens and RoPE. Our current checkpoint starts from the 5,000-step native Landmark64 model, followed by 200 steps of answer-plus-EOS adaptation and 200 steps of retaining-head training. It uses eight groups of retention parameters, memory RoPE, and 128 exact-cache tokens (64 local plus 64 historical).

Table 5: LongBench results grouped by task family (0–100; higher is better). The bottom row reports the equal-weight mean over the seven evaluated tasks. Dashes denote unavailable scores.

Task	LoLCATs	LoLA	Lizard*	Lizard+RH	Liger	GDN	**LYNX**
Single-document QA							
Qasper	1.10	23.10	—	9.98	14.42	6.18	16.01
MultiFieldQA EN	2.50	30.80	—	22.99	21.81	13.30	22.05
Multi-document QA							
HotpotQA	1.00	11.80	19.60	9.70	9.61	7.46	21.96
2WikiMQA	2.10	17.70	16.20	17.04	14.47	9.76	25.60
MuSiQue	0.70	3.40	—	4.76	3.96	4.22	7.73
Few-shot learning							
TREC	8.20	57.50	—	4.50	2.50	26.00	29.50
TriviaQA	1.00	50.20	—	22.13	30.52	68.88	59.63
Seven-task mean	2.37	27.79	—	13.02	13.90	19.40	26.07

Public results retain their source protocols. Lizard* uses an Instruct model; Lizard+RH is our local Base-model adaptation with answer-plus-EOS training and a retaining head. GDN is our local reproduction; Liger denotes Liger-GLA.

Both models use LLaMA-3.1-8B Base and lm-eval 0.4.9.2, greedy decoding, and a 4,096-token input-plus-generation budget. Prompts are left-truncated to retain the suffix; no additional few-shot examples are supplied by the harness, while examples in the task templates remain. Each model is evaluated on 1,350 examples: 150 for MultiFieldQA-en and 200 for each other task. TREC uses classification score and the six QA tasks use F1. All seven tasks use the same checkpoint and yield a mean score of 26.07.

C.1 Extended RULER Protocol

Figure 6 and the (64, 64) row of Table 2 use the same LLaMA-3.1-8B checkpoint as the current LongBench comparison, with eight retention groups, delta updates in both states, and memory RoPE. We evaluate 500 examples per task with greedy decoding and a nominal 4K context budget; actual prompt lengths vary by task. The score is the fraction of reference answers found in the generated text, ignoring case, averaged over examples. Multiple-answer tasks allow partial credit. Generation limits are 128 tokens for retrieval, 30 for variable tracking, 120 for common-word extraction, 50 for frequent-word extraction, and 32 for question answering. MK averages MK1, MK2, and MK3; MQ/MV denote multi-query/multi-value retrieval, VT variable tracking, CWE/FWE common/frequent-word extraction, and HQA/SQA HotpotQA/SQuAD question answering. The displayed Liger-GLA baseline is our local LLaMA-3.1-8B adaptation with 200 answer-plus-EOS steps followed by 200 retaining-head steps. Its cache contains 8 sink, 56 recent, and 64 learned historical tokens. This baseline is distinct from the inference-window-only Liger row in Table 2.

D Ablations and Learned Memory Dynamics

D.1 Multi-Seed General-Task Results

Table 6 reports three 1,000-step LLaMA-3-8B runs with SWA and 16 Landmarks under Appendix B.2's protocol. We average the six tasks within each run, then compute the mean and sample standard deviation across runs.

D.2 Three-Seed MMLU Ablation

Table 7 reports five-shot MMLU accuracy at the 1,000-step checkpoint. We compute the sample standard deviation across three training seeds with denominator $n-1$, where $n=3$. Figure 7 shows one standard deviation across training seeds.

Table 6: General-task accuracy (%) across training seeds. SD uses denominator $n - 1$, with $n = 3$. These statistics supply Table 1.

Task	Seed 1	Seed 2	Seed 3	Mean ± SD
PIQA	79.98	80.96	79.87	80.27 ± 0.60
ARC-e	81.78	81.48	81.23	81.50 ± 0.27
ARC-c	55.55	55.97	54.27	55.26 ± 0.89
HellaSwag	80.31	80.41	80.17	80.30 ± 0.12
WinoGrande	74.74	75.61	75.06	75.14 ± 0.44
MMLU	58.91	57.96	58.00	58.29 ± 0.54
Six-task average	71.88	72.07	71.43	71.79 ± 0.33

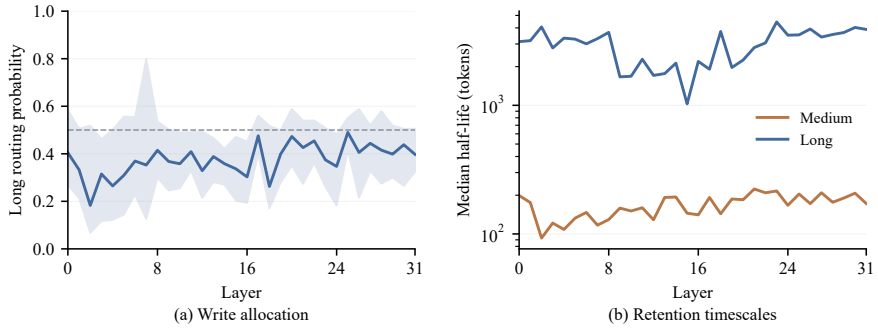
Table 7: MMLU accuracy (%) across three training seeds.

Configuration	Seed 1	Seed 2	Seed 3	Mean ± SD
Single SWA	50.28	53.17	48.49	50.65 ± 2.36
Tied dual states	54.42	51.04	50.45	51.97 ± 2.14
Free dual states	52.41	50.33	54.25	52.33 ± 1.96
Tied + Landmark-8	57.46	57.04	55.39	56.63 ± 1.09
LYNX + Landmark-16	58.91	57.96	58.00	58.29 ± 0.54

D.3 Learned Routing and Retention under a Shared Range

We inspect a separate 5,000-step LLaMA-3.1-8B variant with 16 historical cache tokens. Both states can learn half-lives in $[64, 16384]$, initialized at 256 for Medium and 4,096 for Long. This shared-range setting tests whether distinct retention profiles persist without disjoint bounds. We use two fixed 2,048-token S-NIAH-1 inputs across 32 layers and 32 heads. We denote the Long routing weight by p_L and the Medium weight by $1 - p_L$; zero initialization gives $p_L = 0.5$.

Figure 8 summarizes the layerwise statistics. After training, p_L has mean 0.3922 and standard deviation 0.2079. The pooled median half-lives are 174.2 for Medium and 3,178.8 for Long, estimated from deterministic strided subsamples. Thus, this shared-range variant learns nonuniform write allocation and distinct typical retention timescales. The main model instead uses disjoint ranges.



(a) Write allocation (b) Retention timescales

Figure 8: Learned memory dynamics in the shared-range variant. (a) Long-state routing probability by layer: median and interquartile range over two fixed inputs and their token/head observations; the dashed line marks the initial equal allocation. (b) Median retention half-lives for Medium and Long on a logarithmic axis. Bands summarize variation across the observed states.

E Theoretical Analysis

E.1 Balanced Revisits and the Proof of Lemma 1

Lemma 1 analyzes retention with repeated associative corrections. In this analysis, t counts actual memory-write steps. We use the decay-then-correction recurrence

$$S_t = (I - \beta_t k_t k_t^\top) D_t S_{t-1} + \beta_t k_t v_t^\top. \tag{23}$$

Set $D_t = \alpha I$, $\beta_t = \beta \in (0,1)$, and $S_0 = 0$. Let $U = [u_1, \dots, u_r]$ and $W = [w_1, \dots, w_r]$ satisfy $U^\top U = W^\top W = I_r$, with $2 \le r \le \min(d_k, d_v)$. For $m \ge 2$ rounds, the inputs are

$$k_{(\ell-1)r+i} = u_i, \qquad v_{(\ell-1)r+i} = c_\ell w_i, \qquad c_\ell > 0, \quad 1 \le i \le r, \quad 1 \le \ell \le m. \tag{24}$$

Thus distinct addresses are orthogonal, while the same address is revisited. The value direction is fixed at each address, and its amplitude can vary by round.

Exact state with overwrites. Orthogonality keeps the state in the span of $\{u_i w_i^\top\}_{i=1}^r$. Let $a_{\ell,i}$ be the coefficient of mode i immediately after its update in round ℓ. Between consecutive visits, the mode undergoes r decay steps. Corrections at other addresses leave its direction unchanged, while its own revisit multiplies the decayed coefficient by $1-\beta$ and adds βc_ℓ. Hence

$$a_{\ell,i} = (1-\beta)\alpha^r a_{\ell-1,i} + \beta c_\ell, \qquad a_{0,i} = 0. \tag{25}$$

This recurrence is identical for every address and gives $a_{m,i} = C_m(\alpha,\beta)$ as defined in Equation 4. From the second round onward, the overwrite term $-\alpha\beta u_i u_i^\top S_{t-1}$ is nonzero. At $T = mr$, mode i has undergone $d_i = r - i$ further decay steps, yielding

$$S_T = C_m(\alpha,\beta) \sum_{i=1}^r \alpha^{d_i} u_i w_i^\top = C_m(\alpha,\beta) U \operatorname{diag}(\alpha^{d_1}, \dots, \alpha^{d_r}) W^\top. \tag{26}$$

Proof of Lemma 1. The orthonormal columns of U and W give singular values $C_m(\alpha,\beta)\alpha^{d_i}$. The positive common amplitude cancels in the normalized energies before differentiation, giving $p_i = \alpha^{2d_i} / \sum_j \alpha^{2d_j}$. Set $\theta = -2\log\alpha$, $Z(\theta) = \sum_i e^{-\theta d_i}$, and $\mu(\theta) = \sum_i p_i d_i$, where $p_i = e^{-\theta d_i}/Z$. The spectral entropy is

$$H(p) = -\sum_i p_i \log p_i = \log Z + \theta\mu. \tag{27}$$

Differentiating the normalized weights yields

$$\frac{\mathrm{d}\log Z}{\mathrm{d}\theta} = -\mu, \qquad \frac{\mathrm{d}p_i}{\mathrm{d}\theta} = p_i(\mu - d_i), \qquad \frac{\mathrm{d}\mu}{\mathrm{d}\theta} = -\operatorname{Var}_p(d). \tag{28}$$

Thus $\mathrm{d}H/\mathrm{d}\theta = -\theta \operatorname{Var}_p(d)$. Since $\log r_{\text{eff}} = H$ and $\mathrm{d}\theta/\mathrm{d}\alpha = -2/\alpha$,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \log r_{\text{eff}}(S_T(\alpha)) = \frac{2\theta}{\alpha} \operatorname{Var}_{p(\alpha)}(d). \tag{29}$$

Every weight is positive, and $d_i = r - i$ are distinct for $r \ge 2$. The variance is therefore strictly positive. The derivative is then positive for $0 < \alpha < 1$. Strict monotonicity and continuity at $\alpha = 1$ establish the strict rank comparison in Lemma 1. □

In half-life coordinates, $\alpha = 2^{-1/\tau}$ gives

$$\frac{\mathrm{d}}{\mathrm{d}\tau} \log r_{\text{eff}} = \frac{4(\log 2)^2}{\tau^3} \operatorname{Var}_p(d), \tag{30}$$

which is strictly positive for distinct last-update times and finite $\tau > 0$.

E.2 Spectral Reweighting, a Retention Example, and Robustness

Proof of Lemma 2. The states have equal matrix dimensions and $r \ge 2$ positive singular values each. Their squared singular values are independently sorted, so the ratios $g_i = e_i^A / e_i^B$ compare spectral positions. Normalizing gives

$$p_i^A = \frac{p_i^B g_i}{\sum_j p_j^B g_j}. \tag{31}$$

For a prefix $1, \dots, k$, let $P_k = \sum_{i=1}^k p_i^B$. Because g_i is nondecreasing, its p^B-weighted mean in the prefix is no larger than its mean over all components:

$$\sum_{i=1}^k p_i^A = P_k \frac{\mathbb{E}_{p^B}[g \mid i \le k]}{\mathbb{E}_{p^B}[g]} \le P_k, \qquad 1 \le k < r. \tag{32}$$

Both probability vectors are sorted and have equal total mass. Thus p^B majorizes p^A. Schur concavity of Shannon entropy gives $H(p^A) \geq H(p^B)$ (Li & Busch, 2013). If g is nonconstant, the sorted probability vectors differ, so strict Schur concavity gives $H(p^A) > H(p^B)$. Moreover, $g_1 < \mathbb{E}_{p^B}[g]$ because all $p_i^B > 0$. Equation 31 then gives $p_1^A < p_1^B$. Constant ratios yield identical normalized spectra and equality in both conclusions. □

The criterion concerns the final spectra and permits arbitrary singular-vector bases, access schedules, and gate trajectories that satisfy it. Its connection to a particular retention mechanism requires an update-level argument, as in the following example, or a measurement of the realized energy ratios.

An explicit retention-controlled case. Use the balanced orthogonal revisits of Appendix E.1. Index associations by the number of memory-write steps since their last update, $d_i = i - 1$. For this input construction, value directions $1, \ldots, r_M$ have fixed support in Medium and the remaining directions have fixed support in Long. Require $1 \leq r_M < r$, $r_M \leq d_v/2$, and $r - r_M \leq d_v/2$. The two states together have the same matrix size as Base. Match the inputs, round amplitudes, and effective write coefficient β, where β denotes the coefficient applied after any write allocation. Equation 25 applies within each state. Writing $C_X = C_m(\alpha_X, \beta)$ for $X \in \{B, M, L\}$ yields

$$e_i^B = C_B^2 \alpha_B^{2d_i}, \qquad e_i^A = \begin{cases} C_M^2 \alpha_M^{2d_i}, & i \leq r_M, \\ C_L^2 \alpha_L^{2d_i}, & i > r_M. \end{cases} \tag{33}$$

Here the association index agrees with the spectral order when

$$0 < \alpha_B \leq \alpha_M \leq \alpha_L \leq 1, \qquad C_M \alpha_M^{d_{r_M}} \geq C_L \alpha_L^{d_{r_M+1}}. \tag{34}$$

The second condition preserves the energy order between the two groups. Since $c_\ell > 0$ and $0 < \beta < 1$, $C_m(\alpha, \beta)$ is a polynomial with nonnegative coefficients, hence $C_B \leq C_M \leq C_L$. The relative energy ratios are

$$g_i = \begin{cases} (C_M/C_B)^2 (\alpha_M/\alpha_B)^{2d_i}, & i \leq r_M, \\ (C_L/C_B)^2 (\alpha_L/\alpha_B)^{2d_i}, & i > r_M. \end{cases} \tag{35}$$

Within each group, these ratios are nondecreasing in elapsed time because $\alpha_b/\alpha_B \geq 1$. Across groups, both the amplitude factor and the retention factor are nondecreasing. Thus Equation 15 holds, and Lemma 2 applies. The factors C_M, C_L include the accumulated effects of writing and partial overwriting.

Using $\alpha_b = 2^{-1/\tau_b}$, the ratios become $g_i = (C_{b(i)}/C_B)^2 \, 2^{2d_i(1/\tau_B - 1/\tau_{b(i)})}$, and the boundary condition in Equation 34 is

$$\frac{d_{r_M}}{\tau_M} - \frac{d_{r_M+1}}{\tau_L} \leq \log_2 \frac{C_M}{C_L}, \tag{36}$$

with $1/\tau_b = 0$ when $\alpha_b = 1$. This construction provides an explicit family of retention-controlled updates satisfying the general spectral criterion.

Stability under small perturbations. Let $S^{A,0}, S^{B,0}$ be nonzero reference states with a positive log-effective-rank gap $\Delta = H(p^{A,0}) - H(p^{B,0})$. Let $\widetilde{S}^A, \widetilde{S}^B$ be nonzero perturbed states of the same respective dimensions. Pad their ordered normalized energy spectra with zeros to $n = \min(d_k, d_v) \geq 2$ entries and define

$$\delta_X = \tfrac{1}{2} \|\widetilde{p}^X - p^{X,0}\|_1, \qquad \omega_n(\delta) = \delta \log(n-1) + h_2(\delta), \quad X \in \{A, B\}, \tag{37}$$

where $h_2(\delta)$ is the binary entropy of $(\delta, 1-\delta)$, using natural logarithms. The classical form of the Fannes–Audenaert bound gives $|H(\widetilde{p}^X) - H(p^{X,0})| \leq \omega_n(\delta_X)$ (Audenaert, 2007). Subtracting the two bounds yields

$$\log \frac{r_{\text{eff}}(\widetilde{S}^A)}{r_{\text{eff}}(\widetilde{S}^B)} \geq \Delta - \omega_n(\delta_A) - \omega_n(\delta_B). \tag{38}$$

Consequently, a positive right-hand side preserves the effective-rank advantage even if the perturbed spectra fail the exact ratio ordering. The leading-energy advantage also persists when its reference gap $p_1^{B,0} - p_1^{A,0}$ exceeds $\delta_A + \delta_B$, since each coordinate changes by at most its spectrum's total variation. For a fixed finite input sequence, the complete GDN recurrence and the normalized spectrum of a nonzero state depend continuously on keys, values, and gates. The margin therefore permits sufficiently small nonorthogonal inputs, cross-state value components, and time-varying gate perturbations around the explicit example.

Scope of the retention result. Balanced revisits and spectral ordering are substantive conditions. For unequal revisits with $\beta = 1/2$ and $k_t = v_t$ following (e_1, e_1, e_2), the complete recurrence gives $S_3 = \mathrm{diag}(\alpha/2 + \alpha^2/4, 1/2)$. Its effective rank is 2 at $\alpha = \sqrt{3} - 1$ but approximately 1.8623 at $\alpha = 0.99$. Violating the cross-state energy order can also reverse the comparison: with $r = 4$, $r_M = 2$, $m = 4$, $\beta = 0.4$, $c_\ell = 1$, and $(\alpha_B, \alpha_M, \alpha_L) = (0.95, 0.95, 1)$, the boundary condition in Equation 34 fails. Base has effective rank 3.9739, whereas our two-state construction has 3.9273. For learned inputs and grouped decay, the spectral criterion and perturbation margin can be evaluated on the resulting states.

F Spectral Diagnostics

F.1 Historical Spectral Measurements and Uncertainty

Spectral metrics. We use an energy-based variant of entropy effective rank (Roy & Vetterli, 2007), normalizing squared singular values rather than singular values. For a nonzero state S with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$, define its normalized energy weights and energy-entropy effective rank as

$$p_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2}, \qquad r_{\mathrm{eff}}(S) = \exp\left(- \sum_{i:p_i>0} p_i \log p_i \right). \tag{39}$$

The leading-mode energy fraction and stable rank satisfy

$$p_1 = \frac{\sigma_1^2}{\sum_j \sigma_j^2} = \frac{1}{r_{\mathrm{stable}}(S)}, \qquad r_{\mathrm{stable}}(S) = \frac{\|S\|_F^2}{\|S\|_2^2}. \tag{40}$$

We compute these quantities for each state before aggregating across inputs, layers, and heads.

Historical measurements. Figure 3 compares historical diagnostic checkpoints: our GDN-plus-SWA reproduction and a two-state LYNX variant without memory RoPE. Both streams are fixed RULER NIAH-derived inputs and share some source records. Prefix lengths are 128, 512, 1,024, 2,048, and 4,096. Each model/state-type/length has $2 \times 32 \times 32 = 2048$ observations. Each layerwise summary at 4K contains 64 states from two inputs and 32 heads. Our model's Medium and Long states are concatenated along the value dimension; both models have 128×128 states.

Metrics are computed per state before aggregation. Lines show medians; bands show the 25th–75th percentiles across observed states, with layers and heads nested within each input and model. Measurements at different lengths are terminal states of different prefixes.

Length and layer trends. Across five sampled lengths, our model's median effective rank rises from 9.29 to 15.87. The baseline rises from 5.98 to 8.15 at 2K and then falls to 7.64 at 4K. At 4K, our model's median is approximately $2.08\times$ the baseline's. The two medians reach 12.4% and 6.0% of the maximum rank of 128, respectively. Both states remain spectrally concentrated, with different effective dimensionalities at the same nominal size. The baseline's median p_1 changes from 48.6% at 128 tokens to 32.4% at 4K, while our model's decreases from 31.5% to 17.3%. Our model has a lower leading-mode energy fraction at all measured lengths. Its median effective rank is higher in all 32 layers at 4K. The layer curve preserves the original depth order.

Historical heatmap ordering. For the historical layer/head heatmaps, each cell is the median over the two inputs at 4K. We average the two model matrices with equal weights. Rows and columns are sorted by their descending means in this average. The same permutations are applied to each model. Original IDs and cell values are preserved. Figure 9 compares the original and sorted displays.

State-row cosine similarity. Figures 1 and 4 use terminal states from a separate diagnostic. For state rows $S_{T,i}$ and $S_{T,j}$, each cell is $C_{ij} = \langle S_{T,i}, S_{T,j} \rangle / (\|S_{T,i}\|_2 \|S_{T,j}\|_2)$. All panels use head 0 and the same fixed first DCLM-Edu text prefix at a 2K token budget. Native Gated-DeltaNet-1.3B receives 2,048 tokens and uses zero-based layer 8 with a 256×256 state. The local linearized baseline (*base*) and our historical no-RoPE model use Llama-3.1-8B, receive 1,679 tokens, and use layer 16 with 128×128 states. Our model's two states are concatenated along the value dimension.

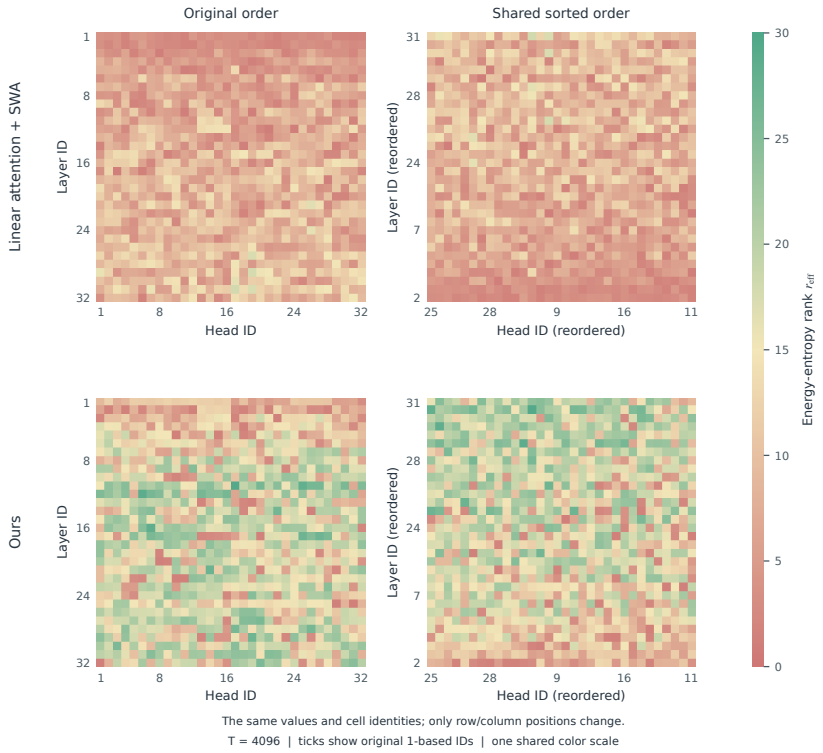


Figure 9: Historical effective-rank heatmaps before and after the shared row/column permutation. Sorting changes display position while preserving cell identities, values, and model-to-model correspondence.

Both axes index the key-address rows of the canonical $[K, V]$ state. The cosine values are uncentered and retain the original row order, with no averaging across inputs or heads. All displayed rows have nonzero norm. The common color scale is $[-1, 1]$. Let D be the diagonal matrix of row norms. The cosine matrix satisfies $S_T S_T^\top = DCD$, whose eigenvalues are the squared singular values used to compute energy-entropy effective rank. Thus row similarities and row magnitudes jointly determine the state's energy spectrum; the heatmap alone omits magnitude differences. Figure 1 annotates normalized effective ranks scaled by 100, $100\, r_{\text{eff}}(S_T)/\min(d_k, d_v)$, computed from the original states, not from C.

Mean absolute off-diagonal cosine is 0.147458, 0.416763, and 0.166748 for native GDN, the baseline, and our model, respectively. The corresponding single-state effective ranks, recomputed with CPU-FP64 SVD, are 41.822954, 10.989304, and 34.219780. These values describe the displayed states. For GDN and Base, normalization by the maximum possible rank gives 0.1634 and 0.0859, respectively. Figure 1 displays these values on a 0–100 scale as 16.34 and 8.59, without a percent sign. The native layer was selected after inspecting earlier layer statistics; head 0 and the first input were fixed without score-based selection. This DCLM example uses a different input protocol from the RULER curves in Figure 3.