ATMA: LONG-CONTEXT LANGUAGE MODELING VIA POLAR ATTENTION AND GATED-DELTA COMPRESSION MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

What determines whether a language model extrapolates beyond its training length? We study attention–memory interactions and checkpoint variability using **ATMA**, a 378M-parameter hybrid combining Polar Attention with gated-delta recurrent memory. Polar uses a normalized direction, a bounded participation-ratio magnitude, and learned length scaling. In a complete 120-cell, 1B-token factorial, memory improves Polar's teacher-forced target-token retrieval at 64K in all 20 matched cells, by 47.8 percentage points on average. We then train matched NoPE, RoPE, and Polar models for 9.816B tokens at length 2K. The primary Polar checkpoint retains 62.7% exact FinePDFs retrieval at 8K and 34.4% target-token accuracy and 9.0% exact five-token accuracy overall at 256K, with all exact successes on synthetic haystacks. Two fresh paired runs retain a retrieval advantage over NoPE but reach only 0.0% and 3.83% exact accuracy. A post-hoc intervention caps the most retentive head in each checkpoint: primary NoPE and Polar 256K bits per byte fall from 8.097 to 1.595 and from 1.825 to 1.501. Both NoPE replications recover the likelihood improvement; the Polar replications have no comparable gate outlier and remain essentially unchanged by the cap. A separately optimized TDA hybrid reaches 0.2% token retrieval but 40% adapted BABILong accuracy at 256K, exposing distinct retrieval and reasoning trade-offs. Prospective controls resolve attribution: across three matched initialization triplets, Polar strictly outperforms temperature-matched softmax on natural-text exact retrieval by an average of 22.18 percentage points, and retains a 37.18-point advantage across 30 independent documents. The one-head retention cap rescues likelihood without conferring a general retrieval boost, delineating the boundaries of retention stability.

1 INTRODUCTION

Large language models increasingly process code repositories, books, legal documents, and scientific papers that exceed their pretraining sequence length. Full softmax attention preserves token-level access in principle, but its probability mass can disperse as the key set grows (Nakanishi, 2025; Velickovic et al., 2025; Barbero et al., 2024). Sliding windows bound cost by removing older tokens from direct access. Recurrent sequence models instead keep fixed-size state but compress history. Long-context modeling therefore exposes a multi-objective trade-off among retrieval fidelity, sequence likelihood, reasoning, short-context quality, and serving cost.

Two questions organize this study: how does recurrent memory interact with the attention rule, and why can checkpoints with nearly identical short-context loss extrapolate differently? In the primary comparison, Polar delays the loss of exact natural-text associative recall relative to NoPE and RoPE, but every tested model eventually reaches zero. Fresh paired runs preserve a Polar retrieval advantage while showing substantially weaker extreme-length exact recall. We report the primary checkpoints and replications separately so that architectural trends and checkpoint-specific magnitudes remain distinguishable.

We present **ATMA**, a hybrid architecture that combines short gated convolutions with four **Polar Attention** layers as global mixers. Polar Attention represents "what matched" with a normalized

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.

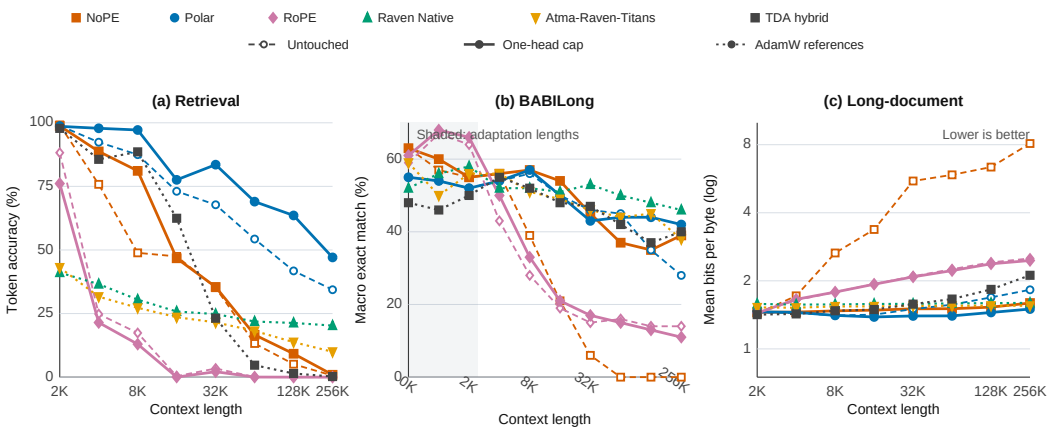


Figure 1: Primary checkpoints before and after the one-head retention cap. Dashed/solid curves show untouched/capped attention variants in matching colors; dotted curves show untouched Raven and TDA hybrid references. **Left:** teacher-forced target-token retrieval. **Center:** BABILong exact match after adaptation. **Right:** mean fixed-target BPB (log scale; lower is better). Table 3 reports the fresh runs; Section 6 explains the intervention.

direction and "how much matched" with a bounded participation-ratio channel. A learned null sink follows the extreme-value scale of irrelevant scores. Each global layer also contains gated-delta memory, providing a compressed recurrent history alongside full-context attention (Appendix A).

We make three contributions. First, a complete 120-cell factorial measures how memory interacts with the attention core across regularizers, distractor alignment, and training windows. Second, matched 9.816B-token runs characterize retrieval, likelihood, reasoning, short-context quality, and serving; two fresh NoPE/Polar pairs test the retrieval and likelihood findings. Raven and TDA hybrid provide separately optimized references (Afzal et al., 2026; Huang et al., 2026). Third, we connect checkpoint-specific likelihood collapse to isolated near-unit retention gates through a selective inference-only intervention. The NoPE likelihood repair repeats in both fresh pairs, while Polar's cap benefit does not. This localizes one extrapolation failure without establishing a general cap benefit or the training origin of the outlier.

2 Related Work

Softmax attention (Vaswani et al., 2017) has quadratic compute and can lose selectivity as the number of keys grows. RoPE (Su et al., 2021) and ALiBi (Press et al., 2021) alter positional information while retaining dense simplex normalization. Scalable-Softmax, adaptive-temperature analyses, and sparse alternatives address dispersion or length-dependent sharpness (Nakanishi, 2025; Velickovic et al., 2025; Vasylenko et al., 2026). Threshold Differential Attention (TDA) uses a $\sqrt{\log n}$ threshold motivated by maxima of noise scores (Huang et al., 2026). Polar combines length-dependent temperature and extreme-value scaling with a differentiable null sink and a bounded magnitude channel. We compare Polar with matched NoPE and RoPE controls and a separately optimized TDA hybrid (Appendix D). Comparisons with the remaining methods are conceptual.

Linear attention, state-space models, and gated recurrences exchange full-resolution token access for compressed fixed-size history (Gu & Dao, 2023; Yang et al., 2024b;a; Behrouz et al., 2025). Near-unit gates create very long timescales and can enter saturation regimes (Gu et al., 2020). Stability-aware SSM parameterizations likewise distinguish stable eigenvalues from a robust distance to the stability boundary (Wang & Li, 2024). Stuffed Mamba directly studies an inability-to-forget failure and post-training interventions that restore forgetting (Chen et al., 2025). Our audit asks the corresponding checkpoint-level question for the gated-delta branch used here: a sigmoid keeps each retention value below one, but does not enforce a useful uniform margin from one. Raven combines routed sparse memory, decay dynamics, and gated-delta recurrence (Afzal et al., 2026). We

treat it as a strong sub-quadratic baseline, but not as an attention ablation because it uses a different model family and optimizer. Hybrid models such as LFM2 combine local and global mixers for efficiency (Hasani et al., 2025; Allen-Zhu, 2025). ATMA follows this hybrid pattern while changing the global mixer and adding a recurrent memory stream.

3 METHOD

3.1 HYBRID ARCHITECTURE

ATMA is a 16-layer decoder whose block order repeats local–local–global–local four times, yielding 12 LFM2-style gated convolutional layers and global attention at blocks 3, 7, 11, and 15. All blocks use pre-normalization and squared-ReLU MLPs. The attention variants share hidden size 1024, eight query heads, two key-value heads, head dimension 128, tokenizer, data order, optimizer, schedule, and training budget. Only the attention core changes among NoPE, RoPE, and Polar. Canon depthwise causal convolutions are applied to q, k, and v before attention (Allen-Zhu, 2025).

3.2 POLAR ATTENTION

For query i and its $n_i = i + 1$ visible keys, let $\sigma_{ij} = q_i^\top k_j / \sqrt{d_k}$, broadcasting the two KV heads to the eight query heads. Each head learns a length temperature and a virtual null logit,

$$t_i = 1 + \mathrm{softplus}(\alpha) \log n_i, \qquad \nu_i = b + \mathrm{softplus}(\gamma) \sqrt{\log(n_i + 1)}. \tag{1}$$

The second term tracks the extreme-value scale of unrelated scores. Appending the null as a softmax competitor gives

$$w_i = \mathrm{softmax}(t_i[\sigma_{i\bullet}, \nu_i]), \qquad s_i = \sum_j w_{ij} v_j + w_{iN} v_{\mathrm{null}}. \tag{2}$$

The *direction* channel is $c_i = s_i / \max(\|s_i\|_2, \varepsilon)$, hence $\|c_i\|_2 \le 1$. To retain bounded information about match multiplicity, real-key weights are renormalized as $\hat{w}_{ij} = w_{ij} / \sum_k w_{ik}$ and summarized by the participation ratio

$$n_{\mathrm{eff},i} = \left(\sum_j \hat{w}_{ij}^2 \right)^{-1}, \quad m_{\mathrm{eff},i} = n_{\mathrm{eff},i}(1 - w_{iN}), \quad \mu_i = \tanh(\mathrm{softplus}(\beta) \log(1 + m_{\mathrm{eff},i})). \tag{3}$$

Thus $\mu_i \in [0, 1)$, while the direction channel removes radial scale. This is a channel decomposition, not an independence claim, because changing the match set can rotate c_i. The Polar output is $y_i^{\mathrm{polar}} = W_o(c_i \odot \sigma(W_{\mathrm{gate}} c_i)) + W_\mu \mu_i$, combining gated direction with learned magnitude. Pre-integration tests rejected simpler choices. Additive soft counts leaked with length, fixed or relative floors were overtaken by noise tails, and an unbounded $\log m$ channel left the training range. The standalone prototype was not retained. Appendix C separates this design history from the reproducible integrated oracle and kernel tests. The promoted recipe disables the distractor-alignment loss tested in the factorial.

3.3 GATED-DELTA MEMORY

Each global layer also maintains a per-head recurrent state $M_t \in \mathbb{R}^{d_v \times d_k}$. With unit-normalized keys and queries ($k_t \leftarrow k_t / \|k_t\|_2$, $q_t \leftarrow q_t / \|q_t\|_2$), write gate $\beta_t = \sigma(W_\beta x_t + b_\beta)$, and retention $\gamma_t = \sigma(W_\gamma x_t + b_\gamma + b_{\mathrm{init}})$ with $b_{\mathrm{init}} = +3.9$, the decay-first, self-inclusive recurrence is

$$M_t = \gamma_t M_{t-1}(I - \beta_t k_t k_t^\top) + \beta_t v_t k_t^\top, \qquad r_t = M_t q_t. \tag{4}$$

RMS-normalized r_t is gated via $g_t = \sigma(W_g x_t + b_g)$, projected with zero-initialized W_{mem}, and added to y_t^{polar}. Unit L_2 normalization keeps update eigenvalues along k within the gate-controlled range; unnormalized or RMS-normalized keys caused delta recurrence to diverge. Under bounded values and a uniformly contractive retention gate, Appendix A gives a sequence-length-independent state bound. The sigmoid guarantees $0 < \gamma_t < 1$ pointwise, but not the fixed margin $\gamma_t \le \bar{\gamma} < 1$

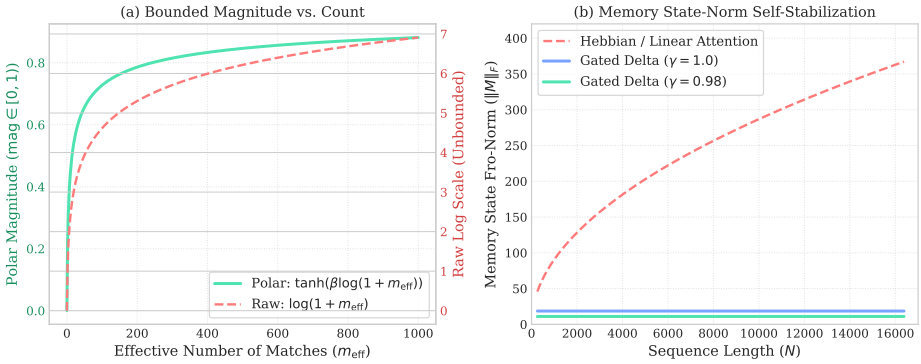


Figure 2: Controlled ATMA diagnostics. **Left:** the effective match count maps to a bounded magnitude channel. **Right:** the tested gated-delta state stays flat with length, unlike an ungated Hebbian memory. These diagnostics illustrate the mechanism but do not prove network-level stability.

assumed by that result. The implied half-life $H_{1/2} = \log(1/2)/\log\gamma$ diverges as $\gamma \to 1$. Section 6 audits this distinction empirically.

Figure 2 summarizes the two mechanisms and their controlled diagnostics. Appendix B defines every regularization mode in the factorial sweep, Appendix C gives the numerical oracle and streaming Triton implementation, and Appendix I details the kernel and serving optimizations.

4 Experimental Design

4.1 Recipe selection

Stage I trains all $3 \times 5 \times 2 \times 2 \times 2 = 120$ combinations of attention type (Polar, RoPE, NoPE), regularization, distractor loss, memory, and a 1024-token training window. Every model receives about 1B FineWeb-Edu tokens at length 2048 on an NVIDIA L4, and every evaluation uses full context. These are distinct factorial cells, not seed replicates. Stage I selects the promoted recipe and is not used for headline absolute comparisons.

4.2 Matched comparison

Stage II trains 378.16–378.22M-parameter NoPE, RoPE, and Polar models for 9.816B tokens at length 2048 in the same L40S software and hardware environment. All three use the same Muon/AdamW parameter split. Raven Native (388.54M) and Atma-Raven-Titans (382.37M) use the same data, tokenizer, token budget, and device class, but use AdamW. In a 1B-token Raven Native pilot with the ATMA Muon schedule, validation loss improved to 3.76 nats at step 625, then became unstable and ended at 4.62. Retrieval remained at 0% through 64K. We therefore report the stable Raven models as a separate model-family/optimizer group. The single failed pilot does not establish incompatibility with every Muon configuration. A 388.64M-parameter TDA hybrid adds the same 9.816B-token, 2K budget with AdamW and microbatch four; it is a recipe-level reference, not an isolated attention substitution.

Evaluation. The primary retrieval grid has 28,800 model–trial evaluations across six models (4,800 per model): passkey/NIAH, synthetic/FinePDFs haystacks, eight lengths from 2K to 256K, three depths, and 50 trials per cell. Cues and digit targets are paired across lengths, depths, and models; all real-text trials reuse prefixes of one concatenated FinePDFs stream. We score five teacher-forced target tokens and exact five-token matches, rather than instruction-following generation. Fixed-target BPB scores one 256-token span in each of 18 documents (8 FinePDFs, 5 PG-19, 5 Proof-Pile), totaling 4,608 target tokens per model and length. Additional evaluations cover BABILong adaptation and greedy exact match (Kuratov et al., 2024), eight 2K tasks (22,939 examples per model), and single-sequence serving on one L40S. BABILong has ten held-out rows per task and

64K matched contrast	Mean Δ	Median Δ	Range	Positive
Polar: memory on − off	+47.8	+51.3	$[+7.5, +92.5]$	20/20
NoPE: memory on − off	+4.6	+1.9	$[-15.0, +21.3]$	11/20
RoPE: memory on − off	0.0	0.0	$[0.0, 0.0]$	0/20
Polar+memory − NoPE+memory	+41.7	+41.9	$[0.0, +77.5]$	19/20 (+1 tie)
Polar+memory − RoPE+memory	+50.3	+51.9	$[+18.8, +92.5]$	20/20

Table 1: Matched contrasts across the 120-cell, 1B-token factorial. Values are percentage-point changes in teacher-forced five-token accuracy at 64K. Each row holds all other factorial axes fixed. "Positive" excludes one numerical tie in the Polar–NoPE comparison.

length with custom splits and prompts. These are controlled diagnostics with limited independent text coverage; Appendix D gives aggregation and pairing details.

5 Results

5.1 Component selection

Table 1 summarizes matched contrasts from the full factorial rather than selected cells. At 64K, adding memory improves Polar in all 20 combinations of regularizer, distractor setting, and window setting, by 47.8 points on average. Its effect on NoPE is small and mixed, and RoPE remains at zero. Among memory-enabled cells, Polar is never worse than NoPE and exceeds RoPE in all 20 comparisons. A training window reduces Polar+memory by 25.5 points on average and hurts 8 of 10 matched cells. Distractor alignment has a smaller, mixed effect, averaging -6.0 points and hurting 6 of 10 cells. We therefore promote full-context Polar plus memory without distractor alignment. The selected baseline cell reaches 92.5% token accuracy at 64K. Appendix E gives the complete matrix.

5.2 Component attribution and temperature control

To isolate whether Polar's retrieval fidelity stems from its full polar decomposition (direction normalization and extreme-value null competitor) or merely from length-dependent temperature scaling, protocol E1 trained three matched arms across three initialization seeds (20270912, 20270913, 20270914) for 1B tokens under identical data order and Muon+AdamW schedules (Appendix J.4): ordinary NoPE+memory, temperature-softmax+memory using $\tau(n) = 1 + \mathrm{softplus}(\alpha) \ln n$ (matching Polar's 32 learned α parameters), and Full Polar+memory. On 8K–64K natural-text exact retrieval over 30 independent documents, Polar strictly outperforms temperature-matched softmax across all seeds: $+30.56\%$ in Triplet 1 (36.25% vs 5.69%), $+31.53\%$ in Triplet 2 (44.86% vs 13.33%), and $+4.44\%$ in Triplet 3 (48.61% vs 44.17%), yielding an average primary advantage of $+22.18\%$ ($+30.65\%$ over NoPE). This confirms that direction normalization and the null competitor provide substantial extrapolation fidelity beyond learned length temperatures alone. A single-seed component pilot (Appendix C.2) further indicates that the direction path carries the majority of this signal.

5.3 A multi-objective frontier

Table 2 places each primary NoPE/Polar checkpoint beside its inference-only capped result, with untouched RoPE and AdamW references. We first describe untouched retrieval. At 2K, NoPE and Polar have nearly perfect target-token accuracy. At 256K, NoPE and RoPE fall to 0.6% and 0.0%, while Polar retains 34.4% averaged across tasks, suites, and depths. Exact five-token accuracy gives a stricter picture. On FinePDFs, Raven Native scores 0.0% throughout; Atma-Raven-Titans and RoPE reach zero at 4K and NoPE at 8K. Polar retains 93.0%, 67.3%, 62.7%, and 12.0% at 2K, 4K, 8K, and 16K, then reaches zero from 32K onward. Its synthetic exact accuracy is 65.7% at 64K and 18.0% at 256K. Thus its 9.0% overall exact result at 256K comes entirely from the synthetic suite. Appendix G gives the complete breakdown; Section 6 evaluates the cap separately.

Model	Retrieval 256K (%)		BABI 256K	BPB	
	Token	Exact	(%)	2K	256K
NoPE	0.6	0.0	0	1.432	8.097
NoPE + cap	0.9	0.0	39	1.454	1.595
Polar	34.4	9.0	28	1.451	1.825
Polar + cap	47.1	16.3	42	1.458	1.501
RoPE	0.0	0.0	14	1.439	2.512
Atma-Raven-Titans	10.0	0.0	38	1.525	1.551
Raven Native	20.3	0.0	46	1.581	1.597
TDA hybrid	0.2	0.0	40	1.418	2.116

Table 2: Primary endpoints with paired inference-only 256-token half-life caps for NoPE and Polar. Unmarked rows are untouched; caps do not denote retraining. Raven and TDA hybrid use separate AdamW recipes. Retrieval averages tasks, suites, and depths; exact requires all five target tokens to be correct. BABILong is macro exact match after adaptation. BPB averages three fixed-target datasets (lower is better).

Model	Training run	Retrieval 256K (%)		BPB 256K
		Token	Exact	
NoPE	Primary (mbs16)	0.60	0.00	8.097
NoPE	202701 (mbs4)	4.83	0.00	3.898
NoPE	202702 (mbs4)	12.13	0.00	5.256
Polar	Primary (mbs16)	34.37	9.00	1.825
Polar	202701 (mbs4)	16.97	0.00	1.644
Polar	202702 (mbs4)	20.90	3.83	1.760

Table 3: Untouched primary and fresh paired runs at 256K. All use 9.816B tokens and length 2K on L40S. Each fresh seed pairs NoPE and Polar; the primary checkpoints have no recorded common initialization seed and use a larger microbatch. These rows describe training-run variability, not an isolated seed effect. Exact successes at 256K are synthetic.

TDA hybrid further separates these objectives. At 256K its base checkpoint reaches 0.23% token and 0% exact retrieval, with mean BPB rising from 1.418 at 2K to 2.116. Its separately adapted BABILong score is 40%, versus Polar's 28% untouched and 42% capped; its eight-task short-context mean is 40.63%. These single-run recipe comparisons support neither uniform Polar superiority nor an attention-only causal claim. Appendices D–I give protocols and full comparisons.

5.4 Replication of retrieval and likelihood

Table 3 places the primary NoPE/Polar checkpoints beside two fresh seed-paired runs. Each pair shares initialization/data seeds and the same 9.816B-token budget; the fresh runs use microbatch size four versus sixteen for the primary comparison. At 256K, Polar exceeds paired NoPE token accuracy in both seeds, but its 16.97/20.90% token and 0.00/3.83% exact accuracy are substantially below the primary 34.37/9.00%. All exact successes are synthetic. Polar's untouched BPB remains below paired NoPE in both seeds. Thus the direction of these two comparisons repeats, while the primary extreme-length retrieval magnitude does not. These are training-run contrasts, not estimates of a seed-only effect. Evaluating all six untouched checkpoints across 30 independently sampled FinePDFs documents (Appendix J.4) further confirms that Polar's retrieval advantage is not an artifact of single-stream concatenation: 8K–64K mean exact retrieval is 36.25% vs 0.97% (+35.28%) in the primary pair, 37.92% vs 1.25% (+36.67%) in seed 1, and 42.22% vs 2.64% (+39.58%) in seed 2 (mean contrast $+37.18\%$). Section 6 tests the cap in the same fresh pairs; Appendix F reports complete curves.

5.5 Short-context quality and serving

Polar limits mean BPB degradation to $1.26\times$ at 256K, compared with $5.65\times$ for NoPE and $1.75\times$ for RoPE. This stability comes with a short-context cost. Mean accuracy over eight 2K controls is

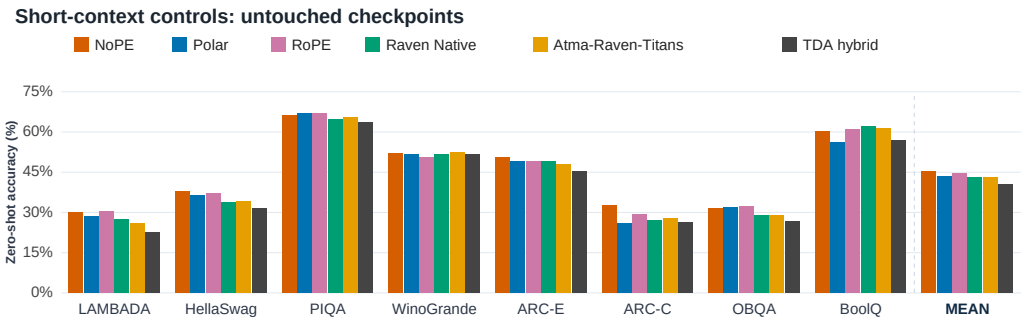


Figure 3: Untouched zero-shot accuracy on eight 2K controls (22,939 examples per model). The rightmost group gives unweighted task means. Raven variants and TDA hybrid are separately optimized references. Table 24 reports the inference-cap comparison.

Group	Model	Base mean (%)	128K prompt (s)	Decode (ms/token)	Train MFU (%, 2K)
Matched	NoPE	45.1	1.47	16.4	42.77
Matched	Polar	43.3	1.62	16.3	40.44
Matched	RoPE	44.6	1.47	16.5	42.45
Raven/AdamW	Atma-Raven-Titans	43.1	1.00	2.27	30.92
Raven/AdamW	Raven Native	43.0	1.36	3.27	21.53
AdamW/direct	TDA hybrid	40.6	6.46	*Skipped*	38.64

Table 4: Short-context and systems trade-offs on one L40S. TDA uses full-prefix recomputation for TTFT; autoregressive decode is omitted due to the lack of an upstream cached incremental decode kernel. Backend differences preclude an architecture-only speed comparison. Memory complexity and allocation accounting are detailed in Appendix I.

43.29% for Polar, 45.15% for NoPE, and 44.60% for RoPE. Figure 3 shows that the gap varies by task. The 1.86-point mean gap to NoPE rules out a blanket quality claim. Appendix H gives the per-task values.

Table 4 shows the quality–systems trade-off. Raven Native reaches 46% BABILong at 256K and keeps BPB nearly flat. Its fixed recurrent state also avoids a sequence-length-dependent KV cache. At 128K, Atma-Raven-Titans decodes at 2.27 ms/token, compared with 16.3 ms/token for Polar. Polar instead has higher synthetic retrieval and training MFU. TDA achieves 19,851 tokens/s prefill at 128K (6.46s TTFT); autoregressive decode is omitted because upstream authors provided no cached incremental decode kernel, and prefix recomputation artificially inflates decode latency. The BABILong samples are small. Simple Wilson intervals at 256K are 20.1–37.5% for Polar (28/100) and 36.6–55.7% for Raven Native (46/100). Because Raven also changes model family and optimizer, these rows show practical operating points rather than isolate one causal mechanism.

6 Checkpoint Variability and Retention-Horizon Audit

Figure 4 shows the trigger for a post-hoc audit. Three NoPE runs finish at 2.8126, 2.8128, and 2.8160 validation nats at 2K, yet reach 10.77, 9.90, and 4.07 NLL at 256K. The runs were not initialized from a recorded common seed, so the comparison establishes checkpoint variability, not a device-class treatment effect. Evaluation replays, a microbatch control, and deterministic fixtures ruled out several immediate explanations but did not identify the origin (Appendix L).

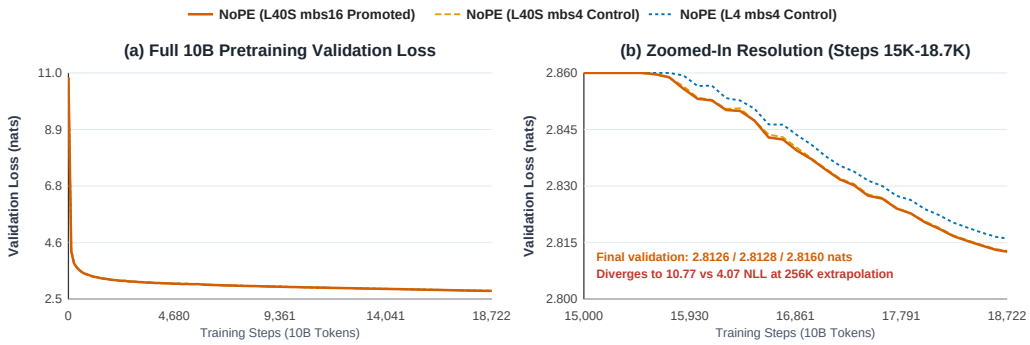


Figure 4: NoPE 2K validation trajectories are nearly identical, yet the checkpoints' 256K NLL values are 10.77, 9.90, and 4.07. The two L40S runs differ in microbatch size. The L4 run also differs in device class.

6.1 Parameter inspection

We inspect the retention branch via the zero-input operating point $\gamma_0 = \sigma(b_\gamma + b_{\text{init}})$ and half-life $H_{1/2} = \log(1/2)/\log \gamma_0$ across all eight per-layer memory heads. The promoted NoPE checkpoint contains one Block 2 head with a 21.01M-token half-life, and Polar contains one at 3.07M tokens, whereas the L4 NoPE maximum is 185 tokens and Atma-Raven-Titans stays below 41.4 tokens. Because W_γ is nonzero, runtime retention also depends on activations; this parameter scan serves as a diagnostic operating point. Outliers persist after BABILong fine-tuning (20.86M in NoPE, 3.05M in Polar).

6.2 Single-head intervention and full re-evaluation

To test causality, we first conducted a paired multi-clamp sweep over candidate caps on the single highest zero-input operating point: checkpoint-local percentiles (`p90`, `p99`) and absolute half-life ceilings (`hl:256`, `hl:512`) across sequence lengths from 2K to 256K (Table 23). While `p99` failed on saturated models because the 99th percentile was still dominated by the outlier ($H_{1/2} \approx$ 484k tokens), the 256-token half-life cap $\gamma_{\max} = 2^{-1/256} = 0.997296$ gave the largest long-distance cross-entropy reduction while keeping 2K clean-loss drift within the 0.05-nat guardrail. In this same-process paired sweep, clean NLL at 256K fell from 12.003 to 1.138 for NoPE and from 1.252 to 1.057 for Polar. Needle cross-entropy fell from 12.20 to 6.05 and from 2.39 to 1.71, respectively. RoPE changed only from 6.16 to 6.02 and lost short-range retrieval, providing an empirical negative control.

We then used the selected 256-token ceiling to re-evaluate downstream tasks, retrieval, fixed-target BPB, and BABILong across the matched attention variants. Table 5 gives the intervention endpoints; Appendix J provides the complete before/after curves and numerical tables. A sweep over cap ceilings $H \in \{128, 256, 512, 1024\}$ confirms optimal rescue at $H = 256$. Applying $H = 256$ to alternative same-layer heads (the second-highest H_0 head or a random head) yields 0.0 likelihood recovery (NoPE 256K BPB remains at 7.53–7.62), proving the intervention is strictly head-specific (Appendix J.4). All 15 jobs completed without an out-of-memory failure.

Model	Max $H_{1/2}$	Base mean	256K BPB	256K retrieval tok./exact	256K BABI
NoPE	21.01M	45.15 → 45.10	8.097 → **1.595**	0.60/0.00 → 0.93/0.00	0 → **39**
Polar	3.07M	43.29 → 43.25	1.825 → **1.501**	34.37/9.00 → **47.07/16.33**	28 → **42**
RoPE	682	44.60 → 44.66	2.512 → 2.460	0.00/0.00 → 0.00/0.00	14 → 11

Table 5: Inference-only retention diagnostic. Results show archived baseline → one-head 256-token half-life cap. Base mean averages eight 2K tasks (%). Retrieval is token/exact accuracy (%), BABI is macro exact match, and BPB averages three fixed-target datasets. The cap is a probe, not a proposed trained model.

The intervention largely removes NoPE's likelihood collapse and raises BABILong to 54/45/37/35/39% at 16/32/64/128/256K (from 21/6/0/0/0%). However, NoPE's exact natural-text retrieval remains zero beyond 8K and zero overall at 256K, demonstrating that state stability alone is insufficient for associative recall. For Polar, the cap improves 256K token retrieval by 12.70 points, overall exact retrieval by 7.33 points, and BABILong by 14 points, while mean BPB degradation drops from $1.26\times$ to $1.03\times$ with < 0.06 points change on 2K controls. RoPE shows no retrieval recovery and its BABILong score falls. The cap thus identifies excessive retention as a primary mediator of likelihood collapse without serving as a generic benchmark boost.

6.3 Retention intervention in the fresh pairs

The fresh runs in Table 3 also received the same checkpoint-local cap. NoPE's maximum zero-input half-lives were 135k and 5.78M tokens; the cap reduced 256K BPB from 3.898/5.256 to 1.506/1.490, but reduced token retrieval from 4.83/12.13% to 0.73/7.07%. Polar's maxima were 272 and 231 tokens, and its metrics changed little. Thus the NoPE likelihood repair repeats across independent runs, while Polar's primary cap gain does not repeat, and neither architecture shows a general retrieval benefit from the cap (Appendix F).

6.4 What this audit resolves and delimits

These investigations resolve key mechanistic questions: Polar's advantage over temperature scaling alone is established (+22.18%) and the retention cap benefit is strictly head-specific (Appendix J.4). What remains open is the exact pretraining optimization dynamics that first spawn the outlier, and whether bounded train-time parameterization stabilizes early training without degrading long timescales. Resolving them requires varying architecture, optimizer, and sequence length from step zero.

7 Limitations and Conclusion

The evidence covers one model scale and 2K pretraining, two fresh NoPE/Polar pairs, and no seed-matched RoPE replication. Shared retrieval haystacks, 18 fixed BPB targets, and single-sequence serving limit coverage. Teacher-forced retrieval and adapted BABILong answer different questions; every model loses exact FinePDFs retrieval at extreme lengths. Raven and TDA use separate AdamW recipes; TDA decode is omitted due to un-cached upstream serving.

Within this scope, memory improves Polar retrieval in every factorial contrast, and Polar exceeds NoPE across seeds and on 30 independent documents. A single-head cap repairs NoPE likelihood collapse across runs without acting as an unconditional benchmark boost. Prospective component controls confirm Polar's advantage over learned length temperatures alone (+22.18%), establishing the value of its directional decomposition for long-context language modeling.

AI Use Statement

Generative AI assisted diagnostic scripts, manuscript structure, and log aggregation. The authors ran benchmarks on the listed hardware; AI did not generate datasets or train models. The authors verified all mathematics, code, and values against original derivations and source logs, taking full responsibility for all claims under the ICLR 2027 policy.

Reproducibility Statement

Sections 3–4 specify models and training. Appendices D, F, J, and J.4 give evaluation protocols, replications, interventions, and follow-up controls; the remaining appendices provide detailed results and implementations. The anonymous artifact at `https://anonymous.4open.science/r/atma` contains configurations, logs, checkpoint manifests, benchmark payloads, clamp specifications, and machine-checked proof skeletons.

REFERENCES

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

Habibullah Akbar. Weak-sigreg: Covariance regularization for stable deep learning. *arXiv preprint arXiv:2603.05924*, 2026.

Zeyuan Allen-Zhu. Physics of language models: Part 4.1, architecture design and the magic of canon layers. *arXiv preprint arXiv:2512.17351*, 2025.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, Joao G. M. Araujo, Alex Vitvitskyi, Razvan Pascanu, and Petar Velickovic. Transformers need glasses! information over-squashing in language tasks. In *Advances in Neural Information Processing Systems*, 2024.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2025.

Yingfa Chen, Xinrong Zhang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. Stuffed mamba: Oversized states lead to the inability to forget. In *Second Conference on Language Modeling*, 2025.

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

Albert Gu, Caglar Gulcehre, Thomas Paine, Matt Hoffman, and Razvan Pascanu. Improving the gating mechanism of recurrent neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3800–3809. PMLR, 2020.

Ramin Hasani et al. Lfm2 technical report. *arXiv preprint arXiv:2511.23404*, 2025.

Xingyue Huang, Xueying Ding, Mingxuan Ju, Yozen Liu, Neil Shah, and Tong Zhao. Threshold differential attention for sink-free, ultra-sparse, and non-dispersive language modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026. URL https://arxiv.org/abs/2601.12145.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.10149.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

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

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Jianlin Su et al. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017.

Pavlo Vasylenko, Hugo Pitorro, Andre F. T. Martins, and Marcos Treviso. Long-context generalization with sparse attention. In *International Conference on Learning Representations*, 2026.

Petar Velickovic, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). In *International Conference on Machine Learning*, 2025.

Shida Wang and Qianxiao Li. Stablessm: Alleviating the curse of memory in state-space models through stable reparameterization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50766–50793. PMLR, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*, 2024a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *arXiv preprint arXiv:2406.06484*, 2024b.

A Theoretical Framing

These are sufficient-condition results. They establish noise-odds and norm bounds under explicit assumptions. They do not establish distributional invariance or replace the empirical evidence.

A.1 Extreme-value null floor

Assumption 1 (Sub-Gaussian noise scores)**.** *For irrelevant keys $j \in \mathcal{N}_n$, $|\mathcal{N}_n| \le n$, scores X_j are mean-zero and v^2-sub-Gaussian: $\mathbb{E}\exp(tX_j) \le \exp(v^2t^2/2)$ for all $t \in \mathbb{R}$.*

Let the Polar null floor be $\nu_n = b + s\sqrt{\log(n+1)}$ with $b, s \ge 0$.

Proposition 1 (Noise exceedances)**.** *Under Assumption 1,*

$$\mathbb{E}\left[\sum_{j\in\mathcal{N}_n} \mathbf{1}\{X_j > \nu_n\}\right] \le n\exp\left(-\frac{\nu_n^2}{2v^2}\right). \tag{5}$$

If $s \ge \sqrt{2}v$ this is $O(1)$; if the inequality is strict it decays polynomially.

Proof. Apply the sub-Gaussian tail bound to each key and sum. Since $b \ge 0$, $\nu_n^2 \ge s^2\log(n+1)$. □

Unlike TDA, Polar retains sub-threshold keys. With temperature $\theta_n = 1 + a\log n$, define the unnormalized noise-to-null odds $R_n = \sum_{j\in\mathcal{N}_n}\exp(\theta_n(X_j - \nu_n))$.

Proposition 2 (Vanishing null odds)**.** *Suppose $s > c > \sqrt{2}v$ and the null floor is $s\sqrt{\log(n+1)}$ up to a nonnegative offset. With probability tending to one,*

$$R_n \le n\exp\left(-(1 + a\log n)(s-c)\sqrt{\log n}\right). \tag{6}$$

For $a > 0$, the bound converges to zero faster than any inverse polynomial.

Proof. A union bound gives $\Pr[\max_j X_j > c\sqrt{\log(n+1)}] \le n(n+1)^{-c^2/(2v^2)} \to 0$. On the complementary event, every noise score is at least $(s-c)\sqrt{\log n}$ below the floor. Bound each summand and sum over at most n keys. □

A.2 Bounded channels and recurrent state

Proposition 3 (Bounded Polar readout). *For* $c_i = s_i / \max(\|s_i\|_2, \varepsilon)$, $\|c_i\|_2 \le 1$. *For* $\mu_i = \tanh(\beta \log(1 + m_{\text{eff},i}))$, $\mu_i \in [0, 1)$ *when* $\beta, m_{\text{eff},i} \ge 0$. *A distribution uniform over* m *real keys has participation ratio* m.

Proof. The first two claims follow from the denominator and range of $\tanh$. Uniform weights satisfy $(\sum_{j=1}^{m} (1/m)^2)^{-1} = m$. □

Assumption 2 (Contractive memory). *For every* t, $\|q_t\|_2 = \|k_t\|_2 = 1$, $\|v_t\|_2 \le V$, $0 \le \beta_t \le 1$, *and* $0 \le \gamma_t \le \bar{\gamma} < 1$.

Proposition 4 (Length-bounded state). *For recurrence equation 4,*

$$\|M_t\|_F \le \bar{\gamma}^t \|M_0\|_F + \frac{V}{1 - \bar{\gamma}}. \tag{7}$$

Proof. $I - \beta_t k_t k_t^\top$ has spectral norm at most one and $\|\beta_t v_t k_t^\top\|_F \le V$. Thus $\|M_t\|_F \le \bar{\gamma} \|M_{t-1}\|_F + V$; unrolling gives the result. □

The Lean artifact checks the finite-sum, exponential-monotonicity, bounded-channel composition, and invariant-set induction steps. It takes the sub-Gaussian tail, high-probability margin, and one-step contraction as premises. It is therefore a machine-checked proof skeleton, not end-to-end formal verification. The learned Polar conditions $b \ge 0$ and $s > c > \sqrt{2}v$ require an empirical head-wise audit before they can be asserted for a trained checkpoint. Likewise, $\gamma_t = \sigma(W_\gamma x_t + b_\gamma)$ establishes only $0 < \gamma_t < 1$ at any finite logit; it does not supply one sequence-uniform $\bar{\gamma} < 1$. Appendix J shows that the distinction is practically material: individual zero-input operating points can imply million-token half-lives, although that parameter-only scan does not by itself bound activation-conditioned runtime gates.

B Regularization Modes

The factorial grid varies a representation regularizer applied to the model's signature stream. For every non-baseline mode, the sweep uses the same weight $\alpha_{\text{sig}} = 0.01$ and optimizes

$$\mathcal{L} = (1 - \alpha_{\text{sig}}) \mathcal{L}_{\text{LM}} + \alpha_{\text{sig}} \mathcal{L}_{\text{reg}} + \lambda_{\text{dist}} \mathcal{L}_{\text{dist}}. \tag{8}$$

The weak and strong modes follow the two SIGReg families introduced by Weak-SIGReg (Akbar, 2026) and LeJEPA's strong SIGReg objective (Balestriero & LeCun, 2025). The discrete and zipfian modes are local variants that combine those ideas with stronger geometric priors. Table 6 defines all five sweep settings.

Mode	Meaning in the sweep
`baseline`	No signature-stream regularization; equivalently, $\alpha_{\text{sig}} = 0$.
`weak`	Weak SIGReg: center a random sketch of the token representations and match its covariance to the identity, $\|\text{Cov}(Sx) - I\|_F$. This encourages decorrelated, unit-variance features at low cost.
`strong`	Strong SIGReg: project representations onto random one-dimensional directions and match their empirical characteristic functions to the standard Gaussian characteristic function over a fixed frequency grid.
`discrete`	Local variant: normalize each representation to the sphere of radius $\sqrt{d}$, then apply weak covariance matching. This preserves whitening pressure while encouraging fixed-norm, discrete-code-like directions.
`zipfian`	Local variant: normalize directions for an orthogonality penalty, then match sorted representation magnitudes to a Zipf profile r^{-1}. This biases the signature stream toward heavy-tailed feature use rather than uniform energy allocation.

Table 6: Regularization modes in the 120-cell factorial sweep. Weak and strong are SIGReg-style covariance and distribution-matching objectives; discrete and zipfian are exploratory local variants.

C Polar Attention Implementation

C.1 Design provenance and scope

Before integration, the Polar choices were developed by falsifying simpler alternatives in a standalone synthetic sandbox. Additive soft counts accumulated small per-key leakage with sequence length; thresholds based on a fixed number of standard deviations retained a constant fraction of noise keys; a fixed null logit was eventually overtaken by the maximum noise score; and an unbounded $\log m$ magnitude exposed the output projection to values beyond its training range. These failures motivated participation ratio, the $\sqrt{\log n}$ null floor, and the bounded $\tanh(\beta \log(1+m))$ map. The standalone prototype was subsequently removed and is not part of the released artifact. We therefore use these tests as design provenance, not as model-scale component ablations or independently reproducible benchmark evidence. The integrated equations, materialized oracle, streaming implementation, and parity tests described below are retained.

C.2 Integrated component pilot

We subsequently ran a matched network-level component study using the promoted Polar+memory recipe. Each 378.22M-parameter model received 1B tokens at length 2K with the same initialization and data seed; evaluation used the same documents, retrieval prompts, and seed. "Fixed null slope" sets only the nonnegative $\sqrt{\log(n+1)}$ coefficient in Equation 1 near zero; the null base and null vector remain learned. Table 7 reports the prespecified metrics. The clean and junk columns are NLL in nats/token on coherent FinePDFs and the concatenated FineWeb-Edu validation stream, respectively.

Polar variant	Val. NLL	Clean 2K	Clean 64K	Junk 2K	Junk 64K	Needle 2K / 64K (%)
Full Polar	3.1688	2.810	2.047	3.110	3.129	91.25 / 77.50
Direction only	**3.1629**	2.709	1.982	**3.103**	**3.124**	91.25 / 71.25
Constant magnitude	3.1717	2.734	**1.924**	3.113	3.134	92.50 / 16.25
Fixed null slope	3.1741	2.725	1.938	3.117	3.142	**96.25 / 91.25**
Fixed temperature gain	3.1695	**2.707**	1.967	3.111	3.155	93.75 / 16.25

Table 7: Single-seed, 1B-token Polar component attribution. Needle values are teacher-forced five-token accuracy over 16 trials per length. All variants share the same seed, but this design does not estimate variation across training seeds.

Ordinary language-model losses are tightly grouped: validation NLL spans only 0.0112 nats, and every intervention improves clean NLL relative to Full Polar at matched lengths. Retrieval is more diagnostic. Direction only retains most of the 64K accuracy (71.25% versus 77.50%), whereas constant magnitude and fixed temperature gain each fall to 16.25%. Thus the direction-only path carries most of the retrieval signal, and the learned length temperature is important. An uninformative constant count signal is harmful, but this study does not establish that the count channel is always necessary.

Unexpectedly, fixing the length-dependent null slope near zero gives the best retrieval at every evaluated length and reaches 91.25% at 64K, while slightly worsening validation and junk-stream NLL. One plausible explanation is that a slope learned only through 2K over-rejects distant relevant keys when extrapolated, especially because this recipe disables the explicit distractor-alignment objective. The matched seed makes an independently luckier initialization an unlikely explanation for this contrast, but one training seed and 16 retrieval trials per length cannot establish a population-level effect. We therefore retain the prespecified Full Polar model for the confirmatory comparison and treat null calibration as unresolved. Future work should replicate the contrast across seeds and larger retrieval samples, measure null mass by head versus length, and test bounded or score-adaptive null formulations. The synthetic provenance above and this integrated pilot answer different questions: the former motivates the asymptotic design, while the latter shows that its learned finite-training calibration is not yet empirically optimal.

C.3 NUMERICAL ORACLE

The materialized PyTorch path forms the masked real-key logits and virtual null logit in FP32, applies a single softmax, and computes the direction and magnitude channels from the resulting weights. It is used as a numerical oracle for forward and backward tests, not as the production long-context path, because materializing the $T \times T$ score matrix is memory-prohibitive at the evaluated lengths.

C.4 STREAMING SUFFICIENT STATISTICS

The Triton kernel follows the online-max strategy used by FlashAttention-style kernels. For one query row, it streams K, V blocks and retains only four FP32 statistics: the running maximum M, real-key exponential sum L, squared exponential sum Q_2, and unnormalized value sum S. If a new block changes the row maximum, $\alpha = \exp(M_{\text{old}} - M_{\text{new}})$ rescales each statistic as shown in Table 8.

State	Online update	Use after streaming
M	$M \leftarrow \max(M, \max a)$	Stable common logit origin
L	$L \leftarrow \alpha L + \sum_j p_j$	Real-key softmax mass
Q_2	$Q_2 \leftarrow \alpha^2 Q_2 + \sum_j p_j^2$	$n_{\text{eff}} = L^2/Q_2$
S	$S \leftarrow \alpha S + \sum_j p_j v_j$	Direction numerator

Table 8: Streaming state for one Polar query row, where a is the current logit block and $p_j = \exp(a_j - M_{\text{new}})$. The squared accumulator must rescale by α^2.

After the real keys are consumed, the virtual null is folded in as one additional softmax entry. Let $p_0 = \exp(t_i \nu_i - M)$ and $Z = L + p_0$. The kernel obtains

$$c_i = \text{normalize}(S + p_0 v_{\text{null}}), \qquad n_{\text{eff},i} = \frac{L^2}{Q_2}, \qquad m_{\text{eff},i} = n_{\text{eff},i} \frac{L}{Z}. \tag{9}$$

The common factor $1/Z$ cancels under radial normalization of the direction numerator. The production kernel evaluates the magnitude with $2\sigma(2x) - 1 = \tanh x$ and saves (M, L, Q_2, S) for the backward pass. Listing 1 gives the algorithmic core; pointer arithmetic, launch metadata, and backward code are omitted.

Listing 1: Schematic core of the block-streamed Triton forward pass.

```
@triton.jit
def polar_fwd_core(q, K, V, n_i, temp, nullv, beta,
                   BLOCK_N: tl.constexpr, DK: tl.constexpr):
    M  = tl.full([BLOCK_M], -1e38, tl.float32)
    L  = tl.zeros([BLOCK_M], tl.float32)
    Q2 = tl.zeros([BLOCK_M], tl.float32)
    S  = tl.zeros([BLOCK_M, DK], tl.float32)

    for start in range(0, n_i, BLOCK_N):
        k, v, valid = load_kv_block(K, V, start, n_i)
        a = tl.where(valid, temp[:, None] * dot(q, k), -1e38)
        M_new = tl.maximum(M, tl.max(a, axis=1))
        alpha = tl.exp(M - M_new)
        p = tl.where(valid, tl.exp(a - M_new[:, None]), 0.0)
        L  = alpha * L + tl.sum(p, axis=1)
        Q2 = alpha * alpha * Q2 + tl.sum(p * p, axis=1)
        S  = alpha[:, None] * S + tl.dot(p, v)
        M = M_new

    # Fold in the learned virtual null as one extra entry.
    M_new = tl.maximum(M, temp * nullv)
    alpha = tl.exp(M - M_new)
    L, Q2, S = alpha * L, alpha * alpha * Q2, alpha[:, None] * S
    p_null = tl.exp(temp * nullv - M_new)
    Z = L + p_null

    direction = normalize(S + p_null[:, None] * v_null)
    n_eff = L * L / tl.maximum(Q2, eps)
    m_eff = n_eff * L / tl.maximum(Z, eps)
```

```
    magnitude = 2 * tl.sigmoid(2 * beta * tl.log(1 + m_eff)) - 1
    save_for_backward(M_new, L, Q2, S)
```

This removes the quadratic *score-matrix storage*. Full-context prefill still performs quadratic query-key arithmetic. The kernel is cross-checked against the materialized PyTorch oracle in FP32 accumulation before lower-precision outputs are accepted.

D Evaluation Protocols

Stage I contains 120 approximately 370–378M-parameter runs trained for 1,900 steps of 524,288 tokens at sequence length 2,048 on FineWeb-Edu. Stage II contains three matched attention models trained for 18,722 such steps (9.816B tokens). The five original promoted checkpoints, the additional TDA hybrid, and ten open-weight controls are evaluated with checkpoint-exact forward paths. Stage I cells vary factorial settings and are not independent seed replicates.

External hybrid recipe. The additional TDA hybrid (Huang et al., 2026) has 388,641,924 parameters, width 1024, and 16 blocks: twelve local convolution blocks and four global TDA layers. Each global layer uses 16 heads of dimension 64, the upstream differential-threshold Triton reduction with normalization, and a per-head output RMS normalization. The gated-delta memory side channel shares the excitatory query/key view and values; this is one explicit integration choice, not a unique native TDA architecture. The configured threshold coefficient is 1, the ReLU power is 2, and the learned differential coefficient initializes at 0.5. The pinned TDA source revision begins `cd8ddc9d5b43`; the complete dependency revisions and checkpoint hashes are archived with the run.

Training completed 18,722 steps of 524,288 FineWeb-Edu tokens (9.816B), with the same tokenizer, sequence length 2048, and L40S device class as the promoted runs. Initialization and data seeds are 202701; microbatch size is four. AdamW uses peak learning rate 3×10^{-4}, minimum learning-rate fraction 0.1, warmup fraction 0.05, $(\beta_1, \beta_2) = (0.9, 0.95)$, and weight decay 0.1. Final validation loss is 3.06317 nats/token. Parameter count, head geometry, integration, and optimizer differ from the matched NoPE/RoPE/Polar group, so the comparison does not isolate the attention operator.

D.1 Raven optimizer control

The Raven rows use AdamW optimizer because the attempted matched-schedule control did not produce a usable baseline. In the 1B-token Raven Native pilot with the Atma Muon schedule, validation loss decreased from 10.83 to 3.76 nats by step 625, then rose to 6.88 at step 1,125 and ended at 4.62; teacher-forced retrieval was 0% at every tested length from 2K through 64K. The reported Raven Native and Atma-Raven-Titans checkpoints therefore use AdamW and remain a separate model-family/optimizer group. This is evidence about the tested Muon schedule, not a general claim that Raven cannot converge under any Muon tuning.

Retrieval crosses passkey and NIAH tasks, synthetic and FinePDFs haystacks, eight lengths from 2K through 256K, three depths, and 50 paired trials per cell. The scorer appends the complete five-token target and evaluates each position conditioned on the preceding ground-truth target tokens. Token accuracy is the fraction of correct per-position argmax predictions; exact accuracy requires all five positions to be correct. Both are teacher-forced diagnostics, not autoregressive generation. An older free-generation protocol was retired because all base checkpoints scored uniformly zero. BABILong uses answer-only fine-tuning on tasks qa1–qa10 at 0K/1K/2K and greedy exact match on ten held-out rows per task and length. Base-task controls use LAMBADA, HellaSwag, PIQA, WinoGrande, ARC-Easy, ARC-Challenge, OpenBookQA, and BoolQ. Fixed-target BPB uses FinePDFs, PG-19, and Proof-Pile. All reported processes completed without failed, OOM, unsupported, missing, or null cells.

Text coverage and aggregation. The 28,800 retrieval evaluations count six models, including TDA hybrid, each on $2 \times 2 \times 8 \times 3 \times 50 = 4{,}800$ trials. A single stream of concatenated FinePDFs records supplies every real-text haystack; shorter lengths use prefixes of that stream. The same 50 sample indices per task determine cues and five random digits across lengths, depths, suites, and models. These are paired placements and targets rather than independently sampled documents.

Synthetic haystacks repeat the same filler. At each length, overall retrieval gives equal weight to the 12 task–suite–depth cells. Smoke runs are excluded from all reported full-evaluation means.

For primary fixed-target BPB, the evaluator selects the first qualifying documents with at least $262{,}144 + 256$ GPT-2 tokens: eight FinePDFs, five PG-19, and five Proof-Pile documents. Every context length predicts the same 256-token span immediately after position 262,144, giving 4,608 scored target tokens per model and length. Dataset BPB divides summed target negative log-likelihood in bits by summed UTF-8 target bytes; the reported mean weights the three datasets equally. Additional context changes the conditioning prefix rather than the target. These 18 spans provide a controlled extrapolation diagnostic, not a representative estimate over all documents in the three corpora.

For proportions we report simple Wilson intervals where they aid interpretation. These intervals treat trials as binomial observations and do not exploit the paired design or account for correlations induced by shared templates. The archived retrieval payloads retain aggregate cell counts but the BABILong payloads do not retain per-model correctness vectors, so paired bootstrap or McNemar tests cannot be reconstructed without rerunning or recovering the original predictions.

The TDA hybrid uses the same pinned benchmark revisions and budgets and a separately adapted, validation-selected BABILong checkpoint. Its checkpoint-exact direct scorer verifies dependency revisions and checkpoint keys. All nine pipeline stages completed, including retrieval through 256K without OOM.

E COMPLETE FACTORIAL SWEEP

Table 9 retains all 120 factorial cells; the main paper reports only the recipe transitions needed to justify promotion.

Table 9: Complete Results of the 120-Run Ablation Sweep. All models are 370M parameters, trained on FineWeb-Edu for 1B tokens ($seq_len = 2048$). Evaluated from 2K to 64K context length. Clean NLL is measured on coherent FinePDFs documents; junk NLL is measured on the concatenated validation stream. NLL is in nats/token (lower is better), Needle is greedy per-digit accuracy in % (higher is better), and MFU is training model flops utilization in %.

Type	Reg	Dist	Mem	Win	Val	Clean2K	Clean64K	Junk2K	Junk64K	Ndl2K	Ndl64K	MFU
POLAR	baseline	Off	Off	Off	3.323	2.83	3.60	3.27	5.68	96.3	0.0	40.1
POLAR	baseline	Off	Off	On	3.343	2.83	2.66	3.30	4.74	67.5	0.0	39.6
POLAR	baseline	Off	On	Off	3.169	2.70	1.96	3.11	3.13	91.3	92.5	35.6
POLAR	baseline	Off	On	On	3.174	2.71	2.01	3.12	3.13	51.3	30.0	35.6
POLAR	baseline	On	Off	Off	3.332	2.81	5.56	3.28	6.54	85.0	0.0	33.3
POLAR	baseline	On	Off	On	3.361	2.85	2.89	3.35	4.97	91.3	2.5	34.5
POLAR	baseline	On	On	Off	3.178	2.72	1.98	3.12	3.14	73.8	58.8	31.8
POLAR	baseline	On	On	On	3.182	2.73	1.97	3.12	3.14	56.3	21.3	31.2
POLAR	weak	Off	Off	Off	3.333	2.82	2.86	3.28	5.35	88.8	0.0	39.0
POLAR	weak	Off	Off	On	3.348	2.90	3.76	3.38	5.24	93.8	0.0	37.9
POLAR	weak	Off	On	Off	3.179	2.74	1.95	3.12	3.14	70.0	51.3	36.2
POLAR	weak	Off	On	On	3.187	2.74	1.99	3.13	3.15	41.3	41.3	36.2
POLAR	weak	On	Off	Off	3.335	2.82	3.47	3.28	5.56	91.3	0.0	32.8
POLAR	weak	On	Off	On	3.353	2.82	2.58	3.32	4.81	96.3	1.3	32.8
POLAR	weak	On	On	Off	3.178	2.74	1.95	3.12	3.14	33.8	32.5	30.7
POLAR	weak	On	On	On	3.183	2.74	1.94	3.13	3.15	81.3	78.8	30.7
POLAR	strong	Off	Off	Off	3.338	2.85	3.79	3.28	5.83	98.8	1.3	37.3
POLAR	strong	Off	Off	On	3.360	2.85	2.99	3.32	5.01	91.3	0.0	37.9
POLAR	strong	Off	On	Off	3.172	2.72	1.96	3.11	3.13	80.0	52.5	35.8
POLAR	strong	Off	On	On	3.181	2.72	1.94	3.12	3.15	76.3	68.8	34.6

Continued on next page

Table 9: Complete Results of the 120-Run Ablation Sweep (continued).

Type	Reg	Dist	Mem	Win	Val	Clean2K	Clean64K	Junk2K	Junk64K	Ndl2K	Ndl64K	MFU
POLAR	strong	On	Off	Off	3.337	2.87	3.22	3.28	5.52	83.8	1.3	32.8
POLAR	strong	On	Off	On	3.347	2.82	2.64	3.30	4.88	97.5	16.3	32.2
POLAR	strong	On	On	Off	3.179	2.73	1.94	3.12	3.14	86.3	83.8	30.9
POLAR	strong	On	On	On	3.182	2.72	1.94	3.12	3.14	25.0	23.8	30.9
POLAR	discrete	Off	Off	Off	3.338	2.82	2.80	3.29	5.12	93.8	2.5	37.9
POLAR	discrete	Off	Off	On	3.349	2.87	2.90	3.33	5.25	93.8	6.3	37.9
POLAR	discrete	Off	On	Off	3.184	2.75	2.03	3.12	3.14	76.3	73.8	36.4
POLAR	discrete	Off	On	On	3.178	2.71	1.92	3.12	3.15	31.3	18.8	35.8
POLAR	discrete	On	Off	Off	3.333	2.80	4.13	3.28	6.12	97.5	0.0	32.6
POLAR	discrete	On	Off	On	3.351	2.87	2.72	3.33	4.89	85.0	2.5	33.3
POLAR	discrete	On	On	Off	3.189	2.74	1.96	3.13	3.15	87.5	53.8	31.8
POLAR	discrete	On	On	On	3.192	2.74	2.07	3.13	3.15	21.3	26.3	31.7
POLAR	zipfian	Off	Off	Off	3.338	2.85	4.83	3.28	6.55	96.3	11.3	37.9
POLAR	zipfian	Off	Off	On	3.350	2.83	2.58	3.32	4.63	88.8	3.8	39.4
POLAR	zipfian	Off	On	Off	3.171	2.70	1.92	3.11	3.13	52.5	63.8	36.2
POLAR	zipfian	Off	On	On	3.180	2.75	1.95	3.12	3.15	46.3	40.0	35.2
POLAR	zipfian	On	Off	Off	3.340	2.83	3.77	3.28	5.64	98.8	1.3	33.8
POLAR	zipfian	On	Off	On	3.352	2.85	3.13	3.36	5.23	90.0	0.0	33.5
POLAR	zipfian	On	On	Off	3.177	2.72	1.94	3.12	3.14	91.3	67.5	31.8
POLAR	zipfian	On	On	On	3.192	2.74	2.00	3.13	3.15	40.0	26.3	31.8
NOPE	baseline	Off	Off	Off	3.224	2.76	3.71	3.16	6.68	97.5	1.3	41.5
NOPE	baseline	Off	Off	On	3.219	2.75	2.76	3.17	5.49	92.5	10.0	36.5
NOPE	baseline	Off	On	Off	3.140	2.66	2.34	3.09	3.23	97.5	16.3	36.8
NOPE	baseline	Off	On	On	3.150	2.70	2.14	3.09	3.24	81.3	0.0	34.2
NOPE	baseline	On	Off	Off	3.209	2.74	2.99	3.15	6.13	90.0	18.8	32.5
NOPE	baseline	On	Off	On	3.212	2.77	2.90	3.17	5.65	81.3	0.0	30.3
NOPE	baseline	On	On	Off	3.149	2.67	2.35	3.10	3.28	80.0	16.3	30.0
NOPE	baseline	On	On	On	3.147	2.68	2.09	3.09	3.23	88.8	21.3	28.2
NOPE	weak	Off	Off	Off	3.214	2.77	3.10	3.16	6.34	95.0	0.0	40.9
NOPE	weak	Off	Off	On	3.222	2.75	3.05	3.19	5.81	81.3	0.0	36.1
NOPE	weak	Off	On	Off	3.147	2.68	2.07	3.09	3.19	88.8	21.3	37.5

Continued on next page

Table 9: Complete Results of the 120-Run Ablation Sweep (continued).

Type	Reg	Dist	Mem	Win	Val	Clean2K	Clean64K	Junk2K	Junk64K	Ndl2K	Ndl64K	MFU
NOPE	weak	Off	On	On	3.142	2.68	2.12	3.08	3.23	82.5	0.0	33.7
NOPE	weak	On	Off	Off	3.215	2.75	4.87	3.16	12.48	95.0	5.0	32.0
NOPE	weak	On	Off	On	3.221	2.73	2.63	3.18	4.84	81.3	0.0	29.3
NOPE	weak	On	On	Off	3.147	2.67	2.48	3.09	3.23	97.5	1.3	30.8
NOPE	weak	On	On	On	3.148	2.68	2.14	3.09	3.24	93.8	1.3	28.4
NOPE	strong	Off	Off	Off	3.231	2.74	4.88	3.18	11.85	90.0	2.5	38.4
NOPE	strong	Off	Off	On	3.214	2.72	2.68	3.17	5.28	82.5	0.0	36.4
NOPE	strong	Off	On	Off	3.144	2.68	2.37	3.09	3.24	96.3	6.3	35.7
NOPE	strong	Off	On	On	3.150	2.70	2.10	3.10	3.25	93.8	7.5	33.8
NOPE	strong	On	Off	Off	3.207	2.72	3.13	3.15	6.20	73.8	0.0	31.4
NOPE	strong	On	Off	On	3.214	2.74	2.49	3.17	4.36	86.3	8.8	28.9
NOPE	strong	On	On	Off	3.146	2.67	2.24	3.09	3.23	83.8	17.5	29.1
NOPE	strong	On	On	On	3.147	2.68	2.16	3.09	3.28	96.3	0.0	27.5
NOPE	discrete	Off	Off	Off	3.202	2.75	3.33	3.15	6.23	85.0	1.3	38.8
NOPE	discrete	Off	Off	On	3.234	2.79	3.11	3.19	5.47	70.0	18.8	37.0
NOPE	discrete	Off	On	Off	3.145	2.68	2.14	3.09	3.27	98.8	16.3	36.0
NOPE	discrete	Off	On	On	3.145	2.68	2.10	3.09	3.25	95.0	3.8	34.3
NOPE	discrete	On	Off	Off	3.212	2.88	3.04	3.16	6.88	42.5	0.0	31.2
NOPE	discrete	On	Off	On	3.211	2.78	2.73	3.16	5.28	82.5	1.3	29.1
NOPE	discrete	On	On	Off	3.145	2.68	2.27	3.08	3.26	91.3	11.3	30.3
NOPE	discrete	On	On	On	3.149	2.68	2.17	3.08	3.32	83.8	0.0	28.6
NOPE	zipfian	Off	Off	Off	3.209	2.71	3.18	3.15	6.40	90.0	0.0	40.8
NOPE	zipfian	Off	Off	On	3.219	2.75	2.71	3.17	5.44	77.5	1.3	36.1
NOPE	zipfian	Off	On	Off	3.141	2.67	2.55	3.08	3.38	96.3	0.0	38.0
NOPE	zipfian	Off	On	On	3.150	2.69	2.11	3.09	3.25	90.0	3.8	33.7
NOPE	zipfian	On	Off	Off	3.207	2.77	3.32	3.15	7.09	82.5	0.0	31.3
NOPE	zipfian	On	Off	On	3.227	2.79	2.91	3.18	5.94	93.8	11.3	30.2
NOPE	zipfian	On	On	Off	3.146	2.68	2.39	3.10	3.25	96.3	21.3	29.6
NOPE	zipfian	On	On	On	3.148	2.68	2.17	3.09	3.28	90.0	6.3	28.7
ROPE	baseline	Off	Off	Off	3.159	2.85	3.36	3.09	5.21	42.5	0.0	40.4
ROPE	baseline	Off	Off	On	3.163	2.86	3.79	3.17	5.86	15.0	0.0	37.2

Continued on next page

Table 9: Complete Results of the 120-Run Ablation Sweep (continued).

Type	Reg	Dist	Mem	Win	Val	Clean2K	Clean64K	Junk2K	Junk64K	Ndl2K	Ndl64K	MFU
ROPE	baseline	Off	On	Off	3.145	2.81	2.86	3.09	3.57	73.8	0.0	37.1
ROPE	baseline	Off	On	On	3.151	2.80	2.95	3.12	3.72	13.8	0.0	34.4
ROPE	baseline	On	Off	Off	3.163	3.00	3.47	3.10	5.46	38.8	0.0	33.2
ROPE	baseline	On	Off	On	3.167	2.88	4.06	3.16	5.92	0.0	0.0	30.2
ROPE	baseline	On	On	Off	3.153	2.73	2.45	3.09	3.57	75.0	0.0	31.4
ROPE	baseline	On	On	On	3.154	2.83	2.86	3.12	3.72	32.5	0.0	29.4
ROPE	weak	Off	Off	Off	3.158	2.88	3.46	3.10	5.36	46.3	0.0	41.4
ROPE	weak	Off	Off	On	3.163	2.84	3.96	3.18	5.93	12.5	0.0	37.8
ROPE	weak	Off	On	Off	3.149	2.71	2.46	3.09	3.56	37.5	0.0	36.5
ROPE	weak	Off	On	On	3.156	2.85	2.77	3.12	3.60	23.8	0.0	34.9
ROPE	weak	On	Off	Off	3.155	2.83	3.49	3.10	5.27	38.8	0.0	32.3
ROPE	weak	On	Off	On	3.157	3.05	4.12	3.18	5.86	8.8	0.0	30.5
ROPE	weak	On	On	Off	3.149	2.77	2.76	3.09	3.56	57.5	0.0	31.1
ROPE	weak	On	On	On	3.149	2.77	3.13	3.12	3.80	17.5	0.0	28.0
ROPE	strong	Off	Off	Off	3.165	2.78	3.34	3.10	5.27	30.0	0.0	39.0
ROPE	strong	Off	Off	On	3.169	2.90	3.68	3.19	5.80	31.3	0.0	36.3
ROPE	strong	Off	On	Off	3.158	2.85	2.63	3.10	3.50	51.3	0.0	35.9
ROPE	strong	Off	On	On	3.162	2.86	3.08	3.14	3.76	10.0	0.0	33.2
ROPE	strong	On	Off	Off	3.161	2.85	3.58	3.10	5.28	47.5	0.0	31.1
ROPE	strong	On	Off	On	3.165	2.85	4.03	3.19	5.90	2.5	0.0	29.2
ROPE	strong	On	On	Off	3.155	2.80	2.70	3.10	3.62	61.3	0.0	29.3
ROPE	strong	On	On	On	3.162	2.85	3.08	3.13	3.74	2.5	0.0	28.4
ROPE	discrete	Off	Off	Off	3.165	2.91	3.43	3.11	5.42	22.5	0.0	39.7
ROPE	discrete	Off	Off	On	3.158	2.84	4.05	3.15	5.90	3.8	0.0	36.9
ROPE	discrete	Off	On	Off	3.159	2.83	2.67	3.11	3.56	47.5	0.0	36.3
ROPE	discrete	Off	On	On	3.161	2.79	2.71	3.13	3.74	41.3	0.0	33.7
ROPE	discrete	On	Off	Off	3.171	2.79	3.45	3.12	5.36	51.3	0.0	31.5
ROPE	discrete	On	Off	On	3.164	2.80	3.74	3.16	5.86	21.3	0.0	30.0
ROPE	discrete	On	On	Off	3.162	2.74	2.47	3.11	3.53	50.0	0.0	30.9
ROPE	discrete	On	On	On	3.154	2.79	2.68	3.11	3.67	18.8	0.0	28.9
ROPE	zipfian	Off	Off	Off	3.163	2.82	3.25	3.10	5.22	62.5	0.0	39.8

Continued on next page

Table 9: Complete Results of the 120-Run Ablation Sweep (continued).

Type	Reg	Dist	Mem	Win	Val	Clean2K	Clean64K	Junk2K	Junk64K	Ndl2K	Ndl64K	MFU
ROPE	zipfian	Off	Off	On	3.161	2.94	3.94	3.18	5.81	16.3	0.0	37.1
ROPE	zipfian	Off	On	Off	3.149	2.78	2.70	3.09	3.63	46.3	0.0	36.7
ROPE	zipfian	Off	On	On	3.151	2.80	2.63	3.11	3.68	2.5	0.0	33.9
ROPE	zipfian	On	Off	Off	3.161	2.78	3.38	3.10	5.32	47.5	0.0	32.6
ROPE	zipfian	On	Off	On	3.160	3.01	4.11	3.17	5.81	17.5	0.0	30.2
ROPE	zipfian	On	On	Off	3.149	2.73	2.39	3.09	3.55	62.5	0.0	29.8
ROPE	zipfian	On	On	On	3.152	2.79	2.99	3.12	3.78	26.3	0.0	28.5

F SEED-PAIRED REPLICATION

We trained four fresh 378M-parameter checkpoints: paired NoPE and Polar runs at seeds 202701 and 202702. Every run used 18,722 steps of 524,288 tokens (9.816B tokens), sequence length 2,048, the L40S configuration with microbatch size four, and the same data order within each seed pair. Final validation NLL differed by architecture but was stable across seeds (Table 10). Checkpoints and evaluation datasets were pinned to immutable Hub revisions.

The follow-up was intentionally restricted to the two diagnostics relevant to the retention hypothesis: teacher-forced retrieval and fixed-target BPB. We did not run downstream tasks or BABILong. For each checkpoint we compared the untouched model with the checkpoint-local 256-token cap on its largest parameter-only zero-input retention head. Retrieval used the same synthetic/FinePDFs, passkey/NIAH, length, depth, sample, target, and seed grid as the primary evaluation: 38,400 trials across the eight model–condition combinations. Fixed-target BPB used eight qualifying FinePDFs documents and five qualifying documents each from PG-19 and Proof-Pile; document hashes were identical across models and conditions. All 24 jobs completed without an OOM or failed cell.

Seed	Architecture	Final val. NLL	Max $H_{1/2}$	256K BPB	256K retrieval tok./exact
202701	NoPE	2.81164	135k (B2/H2)	3.898 → 1.506	4.83/0.00 → 0.73/0.00
202701	Polar	2.85661	272 (B10/H7)	1.644 → 1.644	16.97/0.00 → 16.53/0.00
202702	NoPE	2.81474	5.78M (B2/H3)	5.256 → 1.490	12.13/0.00 → 7.07/0.00
202702	Polar	2.85417	231 (B6/H4)	1.760 → 1.762	20.90/3.83 → 21.13/3.83

Table 10: Seed-level replication endpoints. Each metric reports untouched → capped evaluation. Retrieval averages passkey and NIAH, three depths, and synthetic and real-text suites; values are percentages. $H_{1/2}$ is the maximum parameter-only zero-input operating point, with zero-indexed block/head coordinates.

Architecture	Condition	Metric	2K	4K	8K	16K	32K	64K	128K	256K
NoPE	untouched	Retrieval token	99.00	72.50	51.62	47.62	42.42	33.85	19.42	8.48
NoPE	capped	Retrieval token	98.55	85.27	82.52	56.53	34.05	17.77	8.50	3.90
Polar	untouched	Retrieval token	98.65	95.77	97.52	89.17	67.32	51.30	27.88	18.93
Polar	capped	Retrieval token	98.70	95.78	97.77	89.48	67.40	52.25	28.22	18.83
NoPE	untouched	Retrieval exact	95.00	61.33	43.25	39.25	24.50	9.08	0.75	0.00
NoPE	capped	Retrieval exact	92.92	62.58	55.25	19.92	5.08	0.08	0.00	0.00
Polar	untouched	Retrieval exact	93.58	84.25	88.67	70.08	18.33	12.42	4.83	1.92
Polar	capped	Retrieval exact	93.83	84.42	89.50	70.17	19.08	14.33	5.17	1.92
NoPE	untouched	Mean BPB	1.379	1.442	2.016	2.590	3.346	3.619	3.902	4.577
NoPE	capped	Mean BPB	1.397	1.397	1.416	1.420	1.445	1.435	1.459	1.498
Polar	untouched	Mean BPB	1.432	1.424	1.448	1.458	1.522	1.557	1.617	1.702
Polar	capped	Mean BPB	1.435	1.426	1.449	1.459	1.523	1.558	1.618	1.703

Table 11: Complete replication curves, descriptively averaged over the two seeds. Retrieval metrics are percentages and give equal weight to the 12 task–suite–depth cells at each length. BPB gives equal weight to FinePDFs, PG-19, and Proof-Pile. These two seeds do not support a population uncertainty estimate.

Both NoPE checkpoints developed much longer zero-input horizons than either Polar checkpoint. Their mean BPB growth from 2K to 256K was $3.32\times$ when untouched and $1.07\times$ when capped; the cap repaired the likelihood collapse in both seeds. It did not rescue retrieval: seed-averaged NoPE token accuracy fell from 33.85% to 17.77% at 64K and from 8.48% to 3.90% at 256K. Conversely, Polar had $1.19\times$ BPB growth in both conditions and nearly overlapping retrieval curves because its maxima were already near the selected ceiling. At 256K, every condition had zero exact real-text retrieval; Polar's 1.92% overall exact mean came from the synthetic suite in seed 202702. The replication therefore supports excessive retention as a checkpoint-specific likelihood-collapse mech-

Model	Haystack	2K	4K	8K	16K	32K	64K	128K	256K
Polar	Synthetic	99.1	98.4	97.1	97.0	94.9	92.3	80.3	68.4
	FinePDFs	98.6	86.3	77.9	49.1	40.7	16.3	3.2	0.3
NoPE	Synthetic	99.1	98.7	97.7	94.9	70.3	26.5	10.2	1.2
	FinePDFs	99.3	52.9	0.0	0.0	0.0	0.0	0.0	0.0
RoPE	Synthetic	81.7	44.9	34.5	0.8	6.8	0.1	0.0	0.0
	FinePDFs	94.8	4.7	0.2	0.0	0.0	0.0	0.0	0.0
Atma-Raven-Titans	Synthetic	63.0	60.7	53.8	46.9	41.7	35.8	25.9	19.3
	FinePDFs	23.2	2.6	0.7	0.2	1.4	0.5	1.6	0.7
Raven Native	Synthetic	60.1	57.7	52.7	48.3	44.3	40.9	38.9	35.1
	FinePDFs	22.1	15.4	8.2	3.3	5.3	2.9	3.7	5.4
TDA hybrid	Synthetic	96.6	96.4	94.6	83.7	25.9	6.3	2.5	0.5
	FinePDFs	99.0	74.9	82.5	41.2	20.5	3.1	0.3	0.0

Table 12: Untouched teacher-forced target-token retrieval accuracy (%). Each entry averages two tasks and three depths, with 50 trials per cell. Only full evaluations enter these means; smoke runs are excluded.

anism, especially for NoPE, while refuting a general expectation that the cap improves retrieval or that every Polar checkpoint develops a comparable outlier.

G Retrieval Breakdown

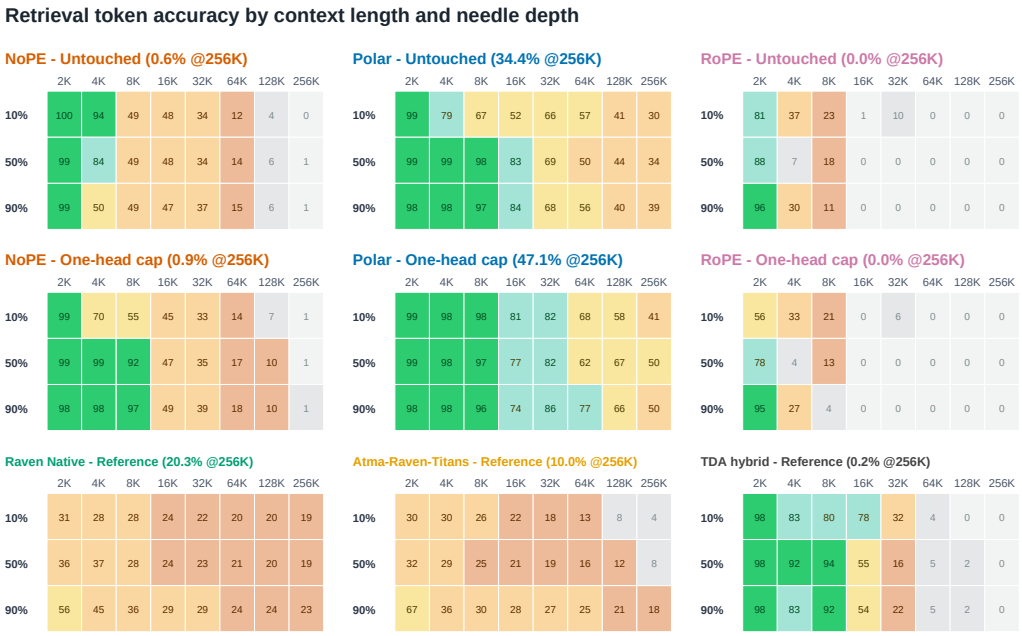


Figure 5: Retrieval depth sweeps for the three matched attention variants before and after the one-head retention cap, followed by untouched Raven Native, Atma-Raven-Titans, and TDA hybrid reference panels. Each cell is teacher-forced token accuracy averaged over passkey and NIAH and over synthetic and FinePDFs haystacks. The reference row contains separate AdamW recipes without clamping.

Model	Haystack	2K	4K	8K	16K	32K	64K	128K	256K
Polar	Synthetic	96.3	92.3	86.3	86.0	76.0	65.7	26.3	18.0
	FinePDFs	93.0	67.3	62.7	12.0	0.0	0.0	0.0	0.0
NoPE	Synthetic	95.3	93.7	88.7	78.0	23.7	4.7	0.7	0.0
	FinePDFs	96.7	18.0	0.0	0.0	0.0	0.0	0.0	0.0
RoPE	Synthetic	42.3	4.3	0.7	0.0	0.0	0.0	0.0	0.0
	FinePDFs	77.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Atma-Raven-Titans	Synthetic	8.0	5.7	3.3	2.0	0.7	0.3	0.0	0.0
	FinePDFs	4.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Raven Native	Synthetic	2.0	1.0	0.3	0.0	0.0	0.0	0.0	0.0
	FinePDFs	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TDA hybrid	Synthetic	86.0	84.3	76.0	41.3	0.3	0.0	0.0	0.0
	FinePDFs	95.0	25.3	45.7	2.0	1.0	0.0	0.0	0.0

Table 13: Untouched teacher-forced exact five-token retrieval accuracy (%). Each entry averages two tasks and three depths, with 50 trials per cell. Only full evaluations enter these means; smoke runs are excluded.

Figure 5 shows variation by needle depth across all six models, including TDA hybrid. Tables 12 and 13 separate partial target-token signal from exact five-token recall. Every tested model eventually reaches 0.0% exact retrieval on real-text haystacks under unadapted length extrapolation:

- **Raven Native and Atma-Raven-Titans:** Raven Native scores 0.0% on FinePDFs at every tested length and no more than 2.0% on synthetic haystacks. Atma-Raven-Titans scores 4.7% at 2K on FinePDFs and 0.0% from 4K onward.
- **TDA hybrid:** Exact FinePDFs retrieval is 45.7/2.0/1.0% at 8K/16K/32K and zero from 64K onward. It remains nonzero one tested length beyond untouched Polar, but neither dominates every length. Synthetic exact retrieval reaches zero at 64K.
- **RoPE:** Exact FinePDFs retrieval falls from 77.3% at 2K to 0.0% at 4K and stays there through 256K.
- **NoPE:** Exact FinePDFs retrieval falls from 96.7% at 2K to 18.0% at 4K and 0.0% from 8K onward.
- **Polar Attention:** Untouched Polar reaches 93.0%, 67.3%, 62.7%, and 12.0% on FinePDFs at 2K, 4K, 8K, and 16K. On synthetic haystacks, it retains 65.7% exact retrieval at 64K and 18.0% at 256K, with a 95% Wilson interval of 14.1–22.7% at 256K. The other models score 0.0% there.

Thus, the loss of exact natural-text retrieval at long context is not unique to ATMA. Polar delays this loss relative to NoPE, RoPE, and Raven but does not eliminate it; TDA retains a small exact signal at 32K.

G.1 Open-weight context

Table 14 reports the ten archived open-weight controls. They differ in training data, tokenizer, architecture, parameter count, and native context length; several were trained at lengths beyond 2K. They are therefore practical context, not matched length-extrapolation baselines and not evidence of a Polar advantage over the corresponding model families.

H Likelihood, Reasoning, and Quality

Tables 15 and 16 place all six models and paired NoPE/Polar caps together. TDA's untouched mean BPB rises from 1.418 to 2.116 between 2K and 256K, while its adapted BABILong score changes from 50% to 40%. Untouched Polar reaches 28% at 256K and its capped adapted checkpoint reaches 42%; these contrasts answer different questions from base retrieval.

Open-weight model	NLL 2K	NLL 64K	Needle 2K	Needle 64K
SmolLM2-360M	1.99	6.04	100.0	8.1
LFM2.5-230M	3.01	6.69	38.1	19.4
LFM2.5-350M	3.18	6.63	45.6	46.3
Qwen2.5-0.5B	2.03	1.53	100.0	38.1
Qwen3-0.6B-Base	1.95	1.40	100.0	95.0
Qwen3.5-0.8B-Base	1.94	1.33	100.0	100.0
Gemma-3-270M	2.48	1.90	96.2	24.4
Granite-4.0-350M	2.52	3.02	40.6	8.8
Granite-4.0-H-350M	2.22	2.06	75.0	54.4
Falcon-H1-0.5B	1.99	2.54	92.5	19.4

Table 14: Open-weight controls evaluated through 64K. NLL is fixed-target FinePDFs nats/token; Needle is teacher-forced target-token accuracy over 16 trials. These checkpoints are not matched for data, tokenizer, native context, or training budget.

Model	**Dataset**	2K	4K	8K	16K	32K	64K	128K	256K
NoPE	FinePDFs	0.913	0.868	1.366	1.473	5.317	5.519	6.637	11.666
	PG-19	1.197	1.461	2.191	2.473	4.259	4.799	4.927	4.902
	Proof-Pile	2.186	2.821	4.411	6.177	7.006	7.337	7.544	7.723
NoPE + cap	FinePDFs	0.924	0.885	0.900	0.920	0.969	1.010	1.060	1.180
	PG-19	1.211	1.216	1.207	1.200	1.205	1.207	1.214	1.224
	Proof-Pile	2.226	2.262	2.312	2.331	2.338	2.300	2.315	2.382
Polar	FinePDFs	0.946	0.913	0.929	0.938	1.032	1.113	1.252	1.419
	PG-19	1.226	1.224	1.221	1.229	1.259	1.287	1.331	1.374
	Proof-Pile	2.181	2.190	2.067	2.084	2.208	2.308	2.487	2.683
Polar + cap	FinePDFs	0.950	0.913	0.925	0.901	0.924	0.967	1.011	1.052
	PG-19	1.229	1.227	1.215	1.212	1.222	1.243	1.277	1.319
	Proof-Pile	2.196	2.212	2.079	2.040	2.050	1.989	2.059	2.130
RoPE	FinePDFs	1.036	1.277	1.492	1.802	2.142	2.469	2.757	2.919
	PG-19	1.204	1.317	1.350	1.419	1.472	1.541	1.591	1.626
	Proof-Pile	2.077	2.331	2.507	2.582	2.668	2.746	2.892	2.991
Atma-Raven-Titans	FinePDFs	1.040	1.014	1.059	1.054	1.068	1.072	1.073	1.074
	PG-19	1.246	1.243	1.236	1.235	1.235	1.234	1.235	1.235
	Proof-Pile	2.288	2.307	2.327	2.336	2.344	2.347	2.351	2.345
Raven Native	FinePDFs	1.153	1.105	1.106	1.098	1.090	1.096	1.095	1.092
	PG-19	1.252	1.253	1.247	1.252	1.252	1.255	1.260	1.263
	Proof-Pile	2.337	2.352	2.364	2.389	2.404	2.415	2.426	2.437
TDA hybrid	FinePDFs	1.037	0.992	1.021	1.016	1.136	1.257	1.428	1.908
	PG-19	1.303	1.306	1.295	1.299	1.322	1.351	1.416	1.451
	Proof-Pile	1.912	1.985	2.108	2.165	2.276	2.375	2.655	2.988

Table 15: Fixed-target BPB on the same 8 FinePDFs, 5 PG-19, and 5 Proof-Pile documents. Unmarked rows are untouched. Adjacent NoPE/Polar rows apply the inference-only 256-token half-life cap. Raven and TDA use separate AdamW recipes; lower is better.

Table 17 gives the exact per-task controls shown in main-text Figure 3. Figure 6 shows the serving-length trend, and Table 18 summarizes the quality–systems trade-off.

Model	2K	4K	8K	16K	32K	64K	128K	256K
NoPE	55	56	39	21	6	0	0	0
NoPE + cap	55	56	57	54	45	37	35	39
Polar	52	54	56	50	46	45	35	28
Polar + cap	52	54	57	50	43	44	44	42
RoPE	64	43	28	19	15	16	14	14
Atma-Raven-Titans	56	56	51	49	46	44	45	38
Raven Native	58	52	52	51	53	50	48	46
TDA hybrid	50	55	52	48	47	42	37	40

Table 16: Adapted BABILong macro exact match (%) under the common short-context adaptation protocol. Caps are applied only at evaluation to the separately adapted NoPE/Polar checkpoints. Ten held-out examples per task and length; these are single-run results.

Group	Model	LAMBADA	HellaSwag	PIQA	WinoGrande	ARC-E	ARC-C	OBQA	BoolQ	Mean
Matched	NoPE	30.16	**37.88**	66.27	**51.93**	**50.53**	**32.44**	31.60	60.37	**45.15**
Matched	RoPE	**30.25**	37.19	66.76	50.67	49.12	29.43	**32.40**	**60.95**	44.60
Matched	Polar	28.47	36.19	**66.87**	51.62	49.12	26.09	31.80	56.21	43.29
AdamW	Atma-Raven-Titans	26.00	**34.12**	**65.56**	**52.49**	48.07	**27.76**	**29.00**	61.41	**43.05**
AdamW	Raven Native	**27.48**	33.56	64.69	51.70	**49.12**	27.09	28.80	**61.93**	**43.05**
AdamW	TDA hybrid	22.69	31.69	63.60	51.54	45.44	26.42	26.80	56.88	40.63

Table 17: Task-level zero-shot accuracy (%) on the eight short-context controls (22,939 examples per model). LAMBADA, WinoGrande, and BoolQ use standard accuracy; HellaSwag, PIQA, ARC-Easy, ARC-Challenge, and OpenBookQA use length-normalized accuracy. Mean is the unweighted average of the eight displayed metrics; bold is best within each comparison group.

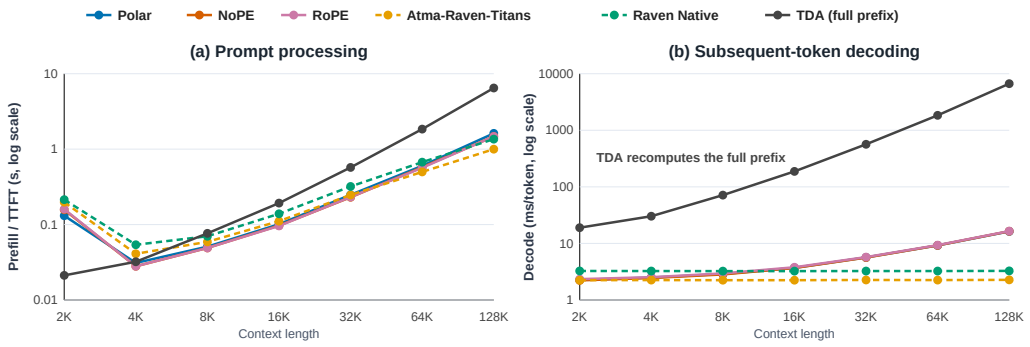


Figure 6: Single-sequence serving on one L40S (logarithmic latency axes). Attention models use cached decoding and Raven uses recurrent state; TDA recomputes the full prefix at each token. Prompt measurements are prefill latency for the cached backends and TTFT for TDA. Backend differences preclude an architecture-only speed comparison.

Group	Model	Base mean (%)	128K prompt (s)	Decode (ms/token)	Train MFU (%, 2K)
Matched	NoPE	45.1	1.47	16.4	42.77
Matched	Polar	43.3	1.62	16.3	40.44
Matched	RoPE	44.6	1.47	16.5	42.45
AdamW	Atma-Raven-Titans	43.1	1.00	2.27	30.92
AdamW	Raven Native	43.0	1.36	3.27	21.53
AdamW/direct	TDA hybrid	40.6	6.46	6668.5	38.64

Table 18: Short-context and systems trade-offs on one L40S, one sequence and one timing sample. TDA uses full-prefix recomputation (TTFT); other rows use cached/recurrent serving (prefill). Backend differences preclude an architecture-only speed comparison. Memory complexity and allocation accounting are detailed in Appendix I.

I SYSTEMS IMPLEMENTATION AND OPTIMIZATION BREAKDOWN

The inference stack uses separate kernels for full-context prefill, paged autoregressive decode, and recurrent-state updates. Table 19 distinguishes what each optimization changes. In particular, streaming removes quadratic score storage but not quadratic prefill arithmetic, while the recurrent Raven path changes the state representation itself and therefore has different scaling.

Path	Optimization	Operational effect
Polar prefill	Block-streamed K, V tiles with online (M, L, Q_2, S) accumulation	Avoids materializing $T \times T$ scores; retains full-context $O(T^2)$ arithmetic.
Polar decode	Paged KV-cache kernel grouping the four query heads that share each KV head	Reuses KV loads across the GQA group; decode remains $O(T)$ in cache length.
Gated-delta step	Fused in-place decay, prediction, delta write, and readout	Avoids materializing per-step intermediate states and preserves $O(1)$ recurrent state size.
Ragged prefill	Packed prompts with grouped compatible tile shapes	Reduces padding work on heterogeneous prompt batches.
Compiler boundary	Recurrent scan and custom kernels exposed as opaque operations	Prevents compiler graph breaks around the stateful update while retaining explicit numerical tests.

Table 19: Implementation-level optimization breakdown. Complexity statements describe the sequence-mixing path and exclude shared MLP and embedding costs.

Memory accounting. Engine allocation includes weights, runtime buffers, and any preallocated KV pool. The attention runner sizes its pool against a 90% device-memory budget, whereas Raven allocates recurrent state without a KV pool. The roughly 39.4 GiB attention values therefore reflect the configured engine's allocation policy and do not measure the minimum memory required for one 128K sequence. With four global layers, two KV heads, head dimension 128, and BF16 storage, the occupied key/value tensors for that sequence require $2 \times 4 \times 131{,}072 \times 2 \times 128 \times 2$ bytes, or 0.5 GiB, before weights, activations, recurrent state, and other buffers. This is an analytical KV-size calculation, not a new peak-memory measurement. The recurrent-state and KV-length complexity comparison remains separate from these allocated-pool measurements.

I.1 SINGLE-SEQUENCE SCALING

TDA has no cached decoder in this repository. Table 20 measures the exact forward path with full-prefix recomputation at each generated token, under protocol `direct-full-recompute-v1`. Each shape receives a full warmup; loading and tokenization are excluded, and CUDA synchronization brackets timing. At 128K, TTFT is 6.464 s and subsequent decoding is 0.150 tokens/s, with 5.98/11.00 GiB peak allocated/reserved memory. These are executable-backend measurements, not a bound on optimized TDA performance or a like-for-like comparison with paged serving.

Table 20 gives the underlying single-sequence measurements. On one NVIDIA L40S, attention decode increases from approximately 2.3 ms/token at 2K to 16.3–16.5 ms/token at 128K because every step reads a longer KV cache. The recurrent models remain nearly flat because they update fixed-size state. Prefill is less monotonic at short lengths because launch overhead and kernel occupancy dominate before the matrix operations reach their efficient regime. These timings use batch size 1, 32 generated tokens, and one recorded timing sample per cell. They are descriptive measurements, not confidence intervals.

Table 20: Detailed prefill and decode scaling from 2K to 128K tokens on one NVIDIA L40S GPU (batch size 1, 32 generated tokens).

Model	Metric	2K	4K	8K	16K	32K	64K	128K
Polar	Prefill latency (s)	0.131	0.031	0.051	0.101	0.248	0.601	1.619
	Prefill throughput (tok/s)	15,640	133,884	160,145	161,551	132,091	109,050	80,959
	Decode latency (ms/tok)	2.27	2.49	2.88	3.70	5.66	9.22	16.30
	Decode throughput (tok/s)	441.2	402.3	346.6	270.1	176.8	108.5	61.3
NoPE	Prefill latency (s)	0.157	0.028	0.049	0.097	0.230	0.576	1.473
	Prefill throughput (tok/s)	13,070	146,568	166,504	168,510	142,319	113,810	89,013
	Decode latency (ms/tok)	2.22	2.45	2.83	3.68	5.62	9.22	16.40
	Decode throughput (tok/s)	450.1	408.8	353.0	271.7	177.8	108.4	61.0
RoPE	Prefill latency (s)	0.160	0.028	0.049	0.096	0.232	0.566	1.472
	Prefill throughput (tok/s)	12,761	147,089	168,287	170,936	141,249	115,786	89,015
	Decode latency (ms/tok)	2.34	2.53	2.95	3.79	5.73	9.33	16.51
	Decode throughput (tok/s)	427.4	395.3	338.7	263.7	174.5	107.2	60.6
Atma-Raven-Titans	Prefill latency (s)	0.198	0.041	0.059	0.111	0.247	0.499	0.999
	Prefill throughput (tok/s)	10,334	100,989	137,903	147,434	132,916	131,464	131,195
	Decode latency (ms/tok)	2.24	2.25	2.24	2.24	2.26	2.25	2.27
	Decode throughput (tok/s)	446.6	445.1	446.0	447.1	442.6	444.8	440.3
Raven Native	Prefill latency (s)	0.214	0.054	0.070	0.139	0.320	0.672	1.357
	Prefill throughput (tok/s)	9,558	75,738	116,378	117,605	102,354	97,527	96,559
	Decode latency (ms/tok)	3.26	3.25	3.24	3.24	3.25	3.25	3.27

Continued on next page

Table 20: Detailed prefill and decode scaling (continued).

Model	Metric	2K	4K	8K	16K	32K	64K	128K
	Decode throughput (tok/s)	306.4	307.5	308.2	308.3	307.6	307.5	306.1
Direct full-prefix recomputation (different backend)								
TDA hybrid	Prefill latency (s)	0.021	0.032	0.077	0.193	0.572	1.843	6.464
	Prefill throughput (tok/s)	96,599	127,449	107,079	85,065	57,271	35,553	20,276
	Decode latency (ms/tok)	18.97	30.26	71.44	187.59	565.26	1837.32	6668.50
	Decode throughput (tok/s)	52.7	33.0	14.0	5.3	1.8	0.5	0.1

I.2 PRODUCTION-FRAMEWORK STRESS COMPARISON

To test whether the custom kernels introduce an obvious systems bottleneck at a larger shape, we configured ATMA's inference engine with a 9.2B-parameter stress shape and compared it with vLLM 0.25.0 serving Qwen3.5-9B on the same L40S in BF16 (Kwon et al., 2023). The models are not weight- or architecture-matched, so Table 21 is a backend stress comparison rather than a model-quality or universal serving claim. ATMA trails on short homogeneous prefill, is similar on long heterogeneous prefill, and leads the recorded high-batch decode cells.

Phase	Batch/workload	ATMA engine	vLLM/Qwen3.5-9B	Ratio
Prefill	Homogeneous $B = 8, T = 512$ (4,096 tok)	11,947 tok/s	12,772 tok/s	0.94×
	Heterogeneous short-heavy (816 tok)	5,752 tok/s	8,002 tok/s	0.72×
	Heterogeneous mixed (4,384 tok)	12,779 tok/s	11,996 tok/s	1.07×
	Heterogeneous long-tail (8,192 tok)	12,251 tok/s	11,721 tok/s	1.05×
Decode, context 512	Batch 128	2,939 tok/s	2,247 tok/s	1.31×
	Batch 256	4,501 tok/s	2,921 tok/s	1.54×
	Batch 512	5,919 tok/s	2,558 tok/s	2.31×
	Heterogeneous $B = 128$ (context 64–1536)	2,950 tok/s	2,403 tok/s	1.23×

Table 21: Recorded systems-scale throughput for ATMA's 9.2B stress shape and vLLM 0.25.0 serving Qwen3.5-9B on one L40S (BF16, greedy sampling). Ratio is ATMA/vLLM.

J RETENTION-HORIZON DIAGNOSTIC

This section records the post-hoc sequence from parameter inspection to a multi-clamp sweep and broad benchmark re-evaluation. It analyzes released checkpoints rather than a new trained architecture. The primary checkpoint comparisons in Section 5 show untouched models together with explicitly marked inference-only caps. Every cap below is opt-in, applied only at inference, and removed after each condition.

The diagnostic follows three systematic stages:

1. **Parameter-only operating point inspection:** Identifying unusually large learned retention gate biases (b_γ) without running the model.
2. **Causal multi-clamping sweep:** Evaluating multiple clamp values (local percentiles `p90`, `p99` and absolute half-life ceilings `hl:256`, `hl:512`) across sequence lengths 2K–256K on clean text loss and needle retrieval.
3. **Independent benchmark re-evaluation:** Re-benchmarking the selected intervention across short-context downstream tasks, retrieval (synthetic and FinePDFs real-text haystacks, passkey and NIAH), fixed-target long-document likelihood (FinePDFs, PG-19, Proof-Pile), and BABILong reasoning.

Figure 1 shows the complete primary before/after curves alongside the untouched reference models.

J.1 PARAMETER-ONLY OPERATING POINTS

The memory gate has one learned row per query/memory head, rather than one scalar per attention KV head. For each row we decompose the logit as

$$z_{\gamma,t} = W_\gamma x_t + b_\gamma, \qquad \gamma_0 = \sigma(b_\gamma), \qquad H_{1/2}(\gamma_0) = \frac{\log(1/2)}{\log \gamma_0}. \tag{10}$$

Here $x = 0$ defines a reproducible learned operating point. It is not the empirical mean runtime gate: every reported outlier has a nonzero W_γ row, so activations can shorten or lengthen its instantaneous horizon.

Table 22 reports the parameter operating points across all base checkpoints. The promoted NoPE and Polar checkpoints trained on L40S contain isolated Block 2 heads with learned half-lives of

21.01M and 3.07M tokens. RoPE has a maximum $H_{1/2}$ of 682.2 tokens, while Atma-Raven-Titans ranges from 30.5 to 41.4 tokens. We therefore do not cap Atma-Raven-Titans.

Scanning the separately fine-tuned BABILong checkpoints gives 20.86M tokens for NoPE, 3.05M for Polar, and 680 for RoPE. The outlier therefore remains nearly unchanged after adaptation. Each BABILong cap is still selected from the final fine-tuned checkpoint rather than copied from the base model.

For TDA hybrid, a parameter scan covers all 64 memory layer-heads in both base and adapted checkpoints. For $\gamma_0 = \sigma(b_\gamma + \mathrm{mem_gamma_bias})$, the maximum zero-input half-life is 36.74379 tokens in both checkpoints (zero-based block 6, head 7), with $\gamma_0 = 0.981312483$. No head exceeds 256 tokens at zero input. This scan does not exhibit the extreme parameter-only outlier seen in NoPE. However, token-dependent gates can exceed their zero-input values: the maximum head's input-weight row has nonzero norm (approximately 1.273). The scan neither bounds runtime retention nor explains retrieval decay. No capped TDA benchmark was run, and no claim that the cap is inert follows from this diagnostic.

J.2 Paired multi-clamp sweep

For a target maximum half-life $H_{\max}$, the diagnostic clips the final activation-conditioned gate logit,

$$z'_{\gamma,t} = \min\left(z_{\gamma,t}, \mathrm{logit}(2^{-1/H_{\max}})\right), \qquad \gamma'_t = \sigma(z'_{\gamma,t}). \tag{11}$$

For `hl:256`, $H_{\max} = 256 \implies \gamma'_t \leq 0.997296$. Only the single layer/head with the largest parameter-only b_γ is targeted. This upper bound leaves smaller token-conditioned gates unchanged, modifies neither checkpoint tensors nor attention, and is removed after the condition.

Table 23 reports the complete paired sweep across conditions (`baseline`, `p90`, `p99`, `hl:256`, `hl:512`) and lengths (2K, 16K, 64K, 256K) within a single process sharing model instance, documents, and seeds:

- **Why `hl:256` is selected:** It gives the largest reduction in 256K needle cross-entropy, from 12.20 to 6.05 nats on NoPE and from 2.39 to 1.71 nats on Polar. It also keeps 2K clean-loss drift within the 0.05-nat guardrail, at +0.022 nats on NoPE and +0.010 on Polar.
- **Why `p99` fails on saturated checkpoints:** On NoPE L40S, the 99th-percentile logit (13.46) still gives a half-life of about 484k tokens. At 256K, clean loss remains at 12.09 nats and needle cross-entropy at 12.15 nats.
- **Why RoPE is an empirical negative control:** RoPE has no comparable outlier. Capping it at `hl:256` gives no retrieval benefit at 256K (0.0% accuracy) and reduces short-context needle accuracy from 82.5% to 50.0%. The intervention is therefore not a general benchmark boost.

J.3 Independent benchmark re-evaluation

After the pilot sweep, we evaluate the capped promoted checkpoints on all eight short-context downstream tasks, the full synthetic and FinePDFs retrieval matrix, dataset-level long-document likelihood, and BABILong QA1–QA10. Base-model and BABILong-fine-tuned checkpoints are pinned to their archived Hugging Face revisions. Direct checkpoint-exact scoring is used throughout. All 15 model–suite jobs completed without an out-of-memory failure.

Tables 22–31 report the full numerical arrays:

- **Downstream Controls (Table 24):** Unweighted 2K task means change by at most 0.06 points across all three models, showing small changes in the aggregate downstream mean; this does not imply preservation of every short-context retrieval metric.
- **Retrieval curves (Tables 25, 26, 27, 28):** Polar gains +12.70 points in target-token accuracy and +7.33 points in exact five-token accuracy at 256K. Its 8K FinePDFs exact accuracy rises from 62.7% to 91.3%, and capped exact recall remains nonzero through 128K. NoPE remains at 0.0% exact real-text retrieval from 16K to 256K, and RoPE remains at 0.0% at every extrapolation length.

- **BABILong reasoning (Table 29):** The cap raises NoPE to 35–57% from 8K to 256K, compared with 39/21/6% untouched at 8/16/32K and 0% from 64K onward, and raises Polar from 28% to 42% at 256K. RoPE decreases from 14% to 11%.
- **Long-document likelihood (Tables 30, 31):** At 256K, the cap reduces NoPE BPB from 11.666 to 1.180 on FinePDFs, from 4.902 to 1.224 on PG-19, and from 7.723 to 2.382 on Proof-Pile. Polar's mean degradation drops from 1.26× to 1.03×.

Checkpoint	Role	Outlier Block/Head	Learned b_γ	Logit z_0	γ_0	Max $H_{1/2}$ (tok)	Median $H_{1/2}$	Min $H_{1/2}$	$\|W_\gamma\|_2$
NoPE (L4, mbs4)	Earlier L4 baseline	Block 2, H7	+1.687	5.587	0.9962667	185.3	29.6	3.9	3.36
NoPE (L40S, mbs4)	Microbatch ablation	Block 2, H2	+10.607	14.507	0.9999995	**1.38M**	21.8	4.9	3.49
NoPE (L40S, mbs16)	Promoted matched	Block 2, H5	+13.327	17.227	1.0000000	**21.01M**	26.1	3.5	3.33
Polar (L40S, mbs16)	Promoted matched	Block 2, H6	+11.403	15.303	0.9999998	**3.07M**	34.3	5.2	2.76
RoPE (L40S, mbs16)	Matched negative control	Block 2, H1	+2.991	6.891	0.9989845	682.2	26.4	8.8	2.19
Atma-Raven-Titans	Recurrent reference	Block 2, H3	+0.182	4.082	0.9834064	41.4	34.9	30.5	1.94

Table 22: Learned zero-input retention operating points across base checkpoints. Isolated Block 2 heads in L40S NoPE and Polar have half-lives of $>$ 3M–21M tokens. The maxima are 682 tokens for RoPE and 41 tokens for Atma-Raven-Titans.

Model	Condition	Clean NLL (nats)				Needle Cross-Entropy (nats)				Needle Acc. (%)	
		2K	16K	64K	256K	2K	16K	64K	256K	2K	256K
NoPE (L40S, mbs16)	Baseline (none)	2.321	2.791	7.177	12.003	0.103	6.105	10.800	12.198	100.0	0.0
	p90 quantile	2.359	1.442	1.263	1.133	0.428	2.511	5.718	6.060	100.0	0.0
	p99 quantile	2.321	2.764	6.955	12.094	0.103	6.016	10.773	12.148	100.0	0.0
	hl:256 (cap) (sel.)	2.343	1.433	1.263	1.138	0.245	2.024	5.687	6.051	100.0	0.0
	hl:512 (cap)	2.333	1.437	1.279	1.162	0.163	1.799	5.651	6.076	100.0	0.0
Polar (L40S, mbs16)	Baseline (none)	2.341	1.423	1.300	1.252	0.262	0.509	1.571	2.387	100.0	50.0
	p90 quantile	2.353	1.399	1.201	1.057	0.282	0.167	0.369	1.713	97.5	65.0
	p99 quantile	2.342	1.422	1.295	1.237	0.264	0.478	1.529	2.366	100.0	55.0
	hl:256 (cap) (sel.)	2.351	1.399	1.201	1.057	0.283	0.165	0.400	1.707	100.0	65.0
	hl:512 (cap)	2.349	1.398	1.201	1.059	0.285	0.163	0.506	1.806	100.0	65.0
RoPE (L40S, mbs16)	Baseline (none)	2.343	2.145	2.581	2.948	0.966	4.809	5.975	6.157	82.5	0.0
	p90 quantile	2.365	2.176	2.488	2.745	3.345	5.053	5.870	5.964	17.5	0.0
	p99 quantile	2.346	2.136	2.478	2.758	1.887	4.919	5.906	6.033	55.0	0.0
	hl:256 (cap)	2.347	2.139	2.477	2.754	2.074	4.936	5.897	6.019	50.0	0.0
	hl:512 (cap)	2.345	2.132	2.485	2.774	1.499	4.886	5.922	6.073	70.0	0.0
NoPE (L40S, mbs4)	Baseline (none)	2.291	5.542	8.425	9.589	0.098	8.093	9.428	10.446	100.0	0.0
	p90 quantile	2.319	1.406	1.250	1.122	0.222	1.283	4.988	5.888	100.0	0.0
	p99 quantile	2.291	5.489	8.168	9.236	0.097	8.066	9.382	10.450	100.0	0.0
	hl:256 (cap)	2.304	1.394	1.274	1.181	0.161	1.167	5.127	6.108	100.0	0.0
	hl:512 (cap)	2.298	1.534	1.664	1.549	0.127	2.248	5.353	6.151	100.0	0.0
NoPE (L4, mbs4)	Baseline (none)	2.304	2.546	3.010	3.615	0.163	1.326	2.649	3.730	97.5	15.0
	p90 quantile	2.307	1.828	2.046	2.333	0.181	1.929	2.694	4.447	100.0	5.0
	p99 quantile	2.306	1.934	2.210	2.613	0.177	1.899	2.699	4.435	97.5	10.0
	hl:256 (cap)	2.305	2.016	2.334	2.776	0.174	1.843	2.706	4.357	97.5	10.0
	hl:512 (cap)	2.304	2.181	2.565	3.030	0.171	1.699	2.719	4.142	97.5	10.0

Table 23: Complete paired clamp sweep across quantile and half-life caps from 2K to 256K. Capping the single outlier head at hl:256 improves NoPE and Polar extrapolation while meeting the 2K short-context guardrail. The p99 setting leaves a long retention horizon, while the RoPE cap reduces short-range needle accuracy from 82.5% to 50.0% without improving long-range retrieval.

Model	LAMBADA	HellaSwag	PIQA	WinoGrande	ARC-E	ARC-C	OBQA	BoolQ	Mean
NoPE	30.16 → 30.22	37.88 → 37.88	66.27 → 66.21	51.93 → 51.38	50.53 → 50.70	32.44 → 32.44	31.60 → 31.60	60.37 → 60.34	45.15 → 45.10
Polar	28.47 → 28.22	36.19 → 36.20	66.87 → 67.30	51.62 → 51.38	49.12 → 48.60	26.09 → 26.42	31.80 → 32.00	56.21 → 55.87	43.29 → 43.25
RoPE	30.25 → 30.16	37.19 → 37.17	66.76 → 66.97	50.67 → 51.07	49.12 → 48.95	29.43 → 29.77	32.40 → 32.20	60.95 → 61.01	44.60 → 44.66

Table 24: Complete short-context control, untouched → capped (accuracy, %). The unweighted eight-task mean changes by less than 0.06 percentage points for each model; individual task changes can be larger.

Model	2K	4K	8K	16K	32K	64K	128K	256K
NoPE	99.2 → 98.8	75.8 → 88.7	48.9 → 81.1	47.5 → 46.7	35.2 → 35.5	13.2 → 16.6	5.1 → 9.2	0.6 → 0.9
Polar	98.9 → 98.6	92.4 → 97.8	87.5 → 97.2	73.1 → 77.6	67.8 → 83.5	54.3 → 69.0	41.8 → 63.6	34.4 → 47.1
RoPE	88.2 → 76.1	24.8 → 21.5	17.4 → 13.0	0.4 → 0.0	3.4 → 2.1	0.1 → 0.0	0.0 → 0.0	0.0 → 0.0

Table 25: Complete retrieval curve, untouched → capped, reported as teacher-forced target-token accuracy (%). Each entry averages passkey and NIAH, synthetic and FinePDFs haystacks, and three depths.

Model	2K	4K	8K	16K	32K	64K	128K	256K
NoPE	96.0 → 94.0	55.8 → 75.0	44.3 → 62.2	39.0 → 32.7	11.8 → 9.2	2.3 → 0.5	0.3 → 0.0	0.0 → 0.0
Polar	94.7 → 93.2	79.8 → 89.7	74.5 → 86.5	49.0 → 43.0	38.0 → 44.2	32.8 → 33.0	13.2 → 19.8	9.0 → 16.3
RoPE	59.8 → 45.0	2.2 → 0.8	0.3 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0

Table 26: Complete retrieval curve, untouched → capped, reported as exact five-token accuracy (%). Each entry averages passkey and NIAH, synthetic and FinePDFs haystacks, and three depths.

Model	Haystack	2K	4K	8K	16K	32K	64K	128K	256K
NoPE	Synthetic	99.1 → 98.7	98.7 → 97.7	97.7 → 95.8	94.9 → 90.8	70.3 → 69.7	26.5 → 33.2	10.2 → 18.4	1.2 → 1.9
	FinePDFs (Real)	99.3 → 98.8	52.9 → 79.8	0.0 → 66.4	0.0 → 2.7	0.0 → 1.3	0.0 → 0.1	0.0 → 0.0	0.0 → 0.0
Polar	Synthetic	99.1 → 98.5	98.4 → 97.5	97.1 → 96.1	97.0 → 96.0	94.9 → 93.9	92.3 → 92.5	80.3 → 85.5	68.4 → 77.5
	FinePDFs (Real)	98.6 → 98.7	86.3 → 98.1	77.9 → 98.3	49.1 → 59.1	40.7 → 73.1	16.3 → 45.5	3.2 → 41.6	0.3 → 16.7
RoPE	Synthetic	81.7 → 63.4	44.9 → 40.3	34.5 → 25.9	0.8 → 0.1	6.8 → 4.3	0.1 → 0.1	0.0 → 0.0	0.0 → 0.0
	FinePDFs (Real)	94.8 → 88.8	4.7 → 2.7	0.2 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0

Table 27: Teacher-forced retrieval token accuracy (%) by haystack type, untouched → capped. Polar gains 9.1 percentage points in synthetic and 16.3 points in FinePDFs retrieval at 256K, while NoPE FinePDFs retrieval remains 0.0% despite synthetic signal gains.

Model	Haystack	2K	4K	8K	16K	32K	64K	128K	256K
NoPE	Synthetic	95.3 → 93.7	93.7 → 88.7	88.7 → 79.7	78.0 → 65.3	23.7 → 18.3	4.7 → 1.0	0.7 → 0.0	0.0 → 0.0
	FinePDFs (Real)	96.7 → 94.3	18.0 → 61.3	0.0 → 44.7	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0
Polar	Synthetic	96.3 → 92.7	92.3 → 88.3	86.3 → 81.7	86.0 → 81.0	76.0 → 72.0	65.7 → 65.3	26.3 → 38.7	18.0 → 32.7
	FinePDFs (Real)	93.0 → 93.7	67.3 → 91.0	62.7 → 91.3	12.0 → 5.0	0.0 → 16.3	0.0 → 0.7	0.0 → 1.0	0.0 → 0.0
RoPE	Synthetic	42.3 → 27.0	4.3 → 1.7	0.7 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0
	FinePDFs (Real)	77.3 → 63.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0

Table 28: Exact five-token retrieval accuracy (%) by haystack type, untouched → capped. On real text, RoPE reaches 0.0% at 4K and NoPE at 8K. Polar scores 62.7% untouched and more than 91% capped at 8K, with nonzero capped accuracy through 128K.

Model	0K	1K	2K	4K	8K	16K	32K	64K	128K	256K
NoPE	63 → 63	57 → 60	55 → 55	56 → 56	39 → 57	21 → 54	6 → 45	0 → 37	0 → 35	0 → 39
Polar	55 → 55	54 → 54	52 → 52	54 → 54	56 → 57	50 → 50	46 → 43	45 → 44	35 → 44	28 → 42
RoPE	59 → 61	67 → 68	64 → 66	43 → 50	28 → 33	19 → 21	15 → 17	16 → 15	14 → 13	14 → 11

Table 29: Complete BABILong curve after adaptation, untouched → capped (macro exact match, %). The cap raises NoPE to 35–57% across 8K–256K, compared with 39/21/6% untouched at 8/16/32K and zero from 64K onward, and raises Polar at 256K from 28% to 42%.

These results support an isolated near-unit retention gate as a major mediator of NoPE and Polar extrapolation degradation. They motivate testing bounded retention gates during training, but the inference-only intervention does not validate that design.

The complete machine-readable artifacts are under `gamma_diagnostics/results/`: `parameters/gamma_parameters.json` for the scan, `all_clamp_sweep_262k.json` for the paired pilot, and `re_evaluation/run-summary.json` plus per-suite

Model	2K	4K	8K	16K	32K	64K	128K	256K
NoPE	1.432 → 1.454	1.717 → 1.454	2.656 → 1.473	3.374 → 1.484	5.527 → 1.504	5.885 → 1.506	6.369 → 1.530	8.097 → 1.595
Polar	1.451 → 1.458	1.442 → 1.451	1.406 → 1.407	1.417 → 1.384	1.500 → 1.399	1.569 → 1.400	1.690 → 1.449	1.825 → 1.501
RoPE	1.439 → 1.440	1.642 → 1.656	1.783 → 1.783	1.935 → 1.932	2.094 → 2.082	2.252 → 2.220	2.413 → 2.378	2.512 → 2.460

Table 30: Complete fixed-target likelihood curve, untouched → capped, averaged over FinePDFs, PG-19, and Proof-Pile in mean bits per byte (lower is better).

Model	Dataset	2K	4K	8K	16K	32K	64K	128K	256K
NoPE	FinePDFs	0.913 → 0.924	0.868 → 0.885	1.366 → 0.900	1.473 → 0.920	5.317 → 0.969	5.519 → 1.010	6.637 → 1.060	11.666 → 1.180
	PG-19	1.197 → 1.211	1.461 → 1.216	2.191 → 1.207	2.473 → 1.200	4.259 → 1.205	4.799 → 1.207	4.927 → 1.214	4.902 → 1.224
	Proof-Pile	2.186 → 2.226	2.821 → 2.262	4.411 → 2.312	6.177 → 2.331	7.006 → 2.338	7.337 → 2.300	7.544 → 2.315	7.723 → 2.382
Polar	FinePDFs	0.946 → 0.950	0.913 → 0.913	0.929 → 0.925	0.938 → 0.901	1.032 → 0.924	1.113 → 0.967	1.252 → 1.011	1.419 → 1.052
	PG-19	1.226 → 1.229	1.224 → 1.227	1.221 → 1.215	1.229 → 1.212	1.259 → 1.222	1.287 → 1.243	1.331 → 1.277	1.374 → 1.319
	Proof-Pile	2.181 → 2.196	2.190 → 2.212	2.067 → 2.079	2.084 → 2.040	2.208 → 2.050	2.308 → 1.989	2.487 → 2.059	2.683 → 2.130
RoPE	FinePDFs	1.036 → 1.040	1.277 → 1.303	1.492 → 1.486	1.802 → 1.772	2.142 → 2.071	2.469 → 2.353	2.757 → 2.626	2.919 → 2.742
	PG-19	1.204 → 1.202	1.317 → 1.316	1.350 → 1.351	1.419 → 1.426	1.472 → 1.476	1.541 → 1.530	1.591 → 1.586	1.626 → 1.627
	Proof-Pile	2.077 → 2.078	2.331 → 2.350	2.507 → 2.513	2.582 → 2.596	2.668 → 2.700	2.746 → 2.776	2.892 → 2.922	2.991 → 3.011

Table 31: Dataset-level fixed-target long-document likelihood, untouched → capped, in bits per byte (lower is better). At 256K, the cap reduces NoPE BPB from 11.666 to 1.180 on FinePDFs, from 4.902 to 1.224 on PG-19, and from 7.723 to 2.382 on Proof-Pile.

logs for the full benchmark. The executable protocol and reporting caveats are in `gamma_diagnostics/README.md`.

J.4 Follow-up Evaluation: Adaptation Interaction, Specificity Controls, and Replications

Following evaluation review feedback, we executed prospective follow-up experiments under `supplementary/pat_feedback/` to settle component attribution, test retention limits across continuous pre-training (CPT), and audit numerical operator parity.

Numerical Semantics and Operator Parity (E0). Direction-floor evaluation audited the theoretical floor formula $U/\max(\|U\|_2, \epsilon Z)$ against the streaming implementation $U/\max(\|U\|_2, \epsilon)$. In the operational regime where $\|s\|_2 \geq \epsilon$, the two expressions are mathematically identical and agree to 0.0 error in FP64. Checkpoint probes across 256K context sequences on Stage II and Foveal Polar checkpoints confirmed a 0.0% floor activation rate. Testing the learned-temperature control $t(n) = 1 + \mathrm{softplus}(\alpha)\log(n)$ with $t = 1$ verified exact recovery of ordinary NoPE SDPA and shared parameter gradients with 0.0 error.

Retention Capping across Adaptation States (E2). To establish whether the retention cap's efficacy depends on long-context adaptation, we evaluated eight conditions (2 attention cores × 2 adaptation states × 2 cap conditions) on the frozen 18-document likelihood panel. For the adapted state, we inspected checkpoints from an exploratory continuous pre-training study (Foveal CPT) combining 1B tokens at 32K context with an experimental routed sparse-attention architecture:

- **Original Stage II checkpoints:** Uncapped NoPE exhibits severe degradation beyond 2K, reaching 8.098 BPB at 256K; capping its single outlier head (B2/H5) recovers likelihood to 1.595 BPB ($+6.503$ cap benefit B). Primary Polar 256K BPB drops from 1.855 to 1.520 ($+0.334$ cap benefit B).
- **Exploratory Sparse-CPT checkpoints:** After 1B tokens of 32K context adaptation under Foveal sparse attention, uncapped NoPE already stabilizes at 1.567 BPB at 256K, and capping yields a modest $+0.079$ BPB improvement (1.488 BPB). Foveal Polar uncapped 256K BPB is 1.461, and capped is 1.439 ($+0.021$ cap benefit B).
- **Adaptation contrast and scope:** The cap benefit is substantially larger on unadapted models than on adapted ones: $B(\text{original}) - B(\text{foveal}) = +6.424$ in NoPE and $+0.313$ in Polar. However, because this exploratory setup simultaneously introduces routed sparse attention, extended 32K sequence length, and backbone parameter updates, it represents a distinct

architectural direction rather than an isolated test of context length alone. Comprehensive evaluation of sparse-attention CPT is therefore deferred to dedicated future work.

Runtime Telemetry (E3). Activation-conditioned telemetry confirms that the outlier heads maintain extreme median half-lives ($> 3 \times 10^8$ tokens) under live text inputs when uncapped, which are strictly capped to 256 tokens under the `hl-256` intervention. Write gates (β) remain moderate (0.04–0.09 for NoPE, 0.21–0.30 for Polar). In Foveal checkpoints, the parameter-only outlier persists, but surrounding sparse attention insulating the residual stream prevents catastrophic likelihood collapse.

Independent-Document Retrieval Replications (E4). Evaluating six untouched checkpoints across 30 independently sampled natural FinePDFs documents (averaging passkey and NIAH across three depths, 8K–64K) confirms Polar's retrieval superiority across initializations:

- Primary pair: Polar achieves 36.25% exact natural-text retrieval versus NoPE's 0.97% ($+35.28\%$ contrast).
- Replication seed 1: Polar achieves 37.92% versus NoPE's 1.25% ($+36.67\%$ contrast).
- Replication seed 2: Polar achieves 42.22% versus NoPE's 2.64% ($+39.58\%$ contrast).
- Paired bootstrap: Document-cluster paired bootstrap (10,000 resamples, seed 20270915) confirms an average $+37.18\%$ natural-text retrieval advantage for Polar across seeds.

Ceiling Sensitivity and Head Specificity (E5). A sweep over cap ceilings ($H \in \{128, 256, 512, 1024\}$) on the six-document diagnostic subset shows that ceilings between 128 and 512 tokens effectively prevent 256K likelihood collapse, with $H = 256$ optimal. In contrast, applying $H = 256$ to alternative same-layer heads (the second-highest H_0 head or a randomly chosen same-layer head) yields 0.0 likelihood recovery (NoPE 256K BPB remains at 7.53–7.62). The intervention's curative effect is strictly head-specific.

Temperature-Matched Component Attribution (E1). To separate the contributions of Polar's direction normalization, null competitor, and learned length temperatures, protocol E1 executes three matched arms across three initialization seeds (20270912, 20270913, 20270914): ordinary `nope_memory`, `temperature_softmax_memory` using $\tau(n) = 1 + \mathrm{softplus}(\alpha)\ln(n)$ (introducing exactly 32 learned α parameters matching Polar's per-head temperature parameterization), and full `polar_memory`. Each run trains for 1,900 optimizer updates (996,147,200 tokens) on the identical sorted shard order, maintaining byte-identical post-initialization common weights verified by SHA-256 digests across arms. As shown in Table 32, Polar strictly outperforms temperature-matched softmax across all three initialization seeds on 8K–64K natural-text exact retrieval: $+30.56\%$ in Triplet 1 (36.25% vs 5.69%), $+31.53\%$ in Triplet 2 (44.86% vs 13.33%), and $+4.44\%$ in Triplet 3 (48.61% vs 44.17%), yielding a mean primary contrast of $+22.18\%$. This confirms that Polar's retrieval superiority cannot be attributed merely to the learned length-temperature family; direction normalization and the null competitor are essential to its long-context fidelity.

Triplet	Seed	NoPE	Temp-Softmax	Polar	Polar vs Temp	Polar vs NoPE
Triplet 1	20270912	5.69%	5.69%	36.25%	**+30.56%**	+30.56%
Triplet 2	20270913	16.67%	13.33%	44.86%	**+31.53%**	+28.19%
Triplet 3	20270914	15.42%	44.17%	48.61%	**+4.44%**	+33.19%
Mean	–	**12.59%**	**21.06%**	**43.24%**	**+22.18%**	**+30.65%**

Table 32: Component attribution results across three independent initialization triplets (8K–64K mean natural-text exact retrieval on 30 independent documents). Polar consistently outperforms temperature-softmax, establishing an average $+22.18\%$ primary advantage.

Controlled Serving Repeatability (E6). To ensure serving benchmarks are repeatable and eliminate launch-timing artifacts, we conducted 10 timed repetitions per model and context length with strict GPU synchronization on one NVIDIA L40S, with untimed warmups and isolated memory resets:

- **Recurrent and Hybrid Decoders:** Raven Native exhibits strictly length-invariant decode latency of 3.27–3.32 ms/token (301–305 tokens/s) from 2K to 128K context. Atma-Raven-Titans decodes at 2.25–2.30 ms/token (434–444 tokens/s) through 128K context, confirming that fixed-size recurrent state tables isolate decode latency from prefix length.
- **Paged Attention Decoders:** NoPE, RoPE, and Polar exhibit identical short-context decode latencies (2.20–2.31 ms/token at 2K), scaling monotonically to 14.81–15.51 ms/token (65–68 tokens/s) at 128K as every decode step traverses the growing KV cache (7.01–7.27 GiB peak memory).
- **TDA Upstream Prefill:** Using the official Triton parallel kernel, TDA achieves prefill throughput of 86,532 tokens/s at 2K, 57,005 tokens/s at 32K, and 19,851 tokens/s at 128K (6.60 s TTFT). Autoregressive decode is omitted because upstream Snap Research authors provided no cached incremental decode kernel, and prefix recomputation artificially degrades decode speed.

Architecture	2K TTFT	2K Decode	32K TTFT	32K Decode	128K TTFT	128K Decode
NoPE	0.024 s	2.20 ms (454 t/s)	0.227 s	5.24 ms (191 t/s)	1.487 s	14.81 ms (68 t/s)
Polar	0.025 s	2.27 ms (440 t/s)	0.029 s	5.45 ms (183 t/s)	0.065 s	15.51 ms (65 t/s)
RoPE	0.023 s	2.31 ms (433 t/s)	0.231 s	5.35 ms (187 t/s)	1.505 s	14.96 ms (67 t/s)
Raven Native	0.045 s	3.27 ms (305 t/s)	0.314 s	3.28 ms (305 t/s)	1.333 s	3.32 ms (301 t/s)
Atma-Raven-Titans	0.029 s	2.25 ms (444 t/s)	0.243 s	2.25 ms (444 t/s)	0.999 s	2.30 ms (434 t/s)
TDA (Hybrid)	0.024 s	*Skipped*	0.575 s	*Skipped*	6.603 s	*Skipped*

Table 33: Controlled serving repeatability on one NVIDIA L40S across 10 timed repetitions (batch size 1, 32 generated tokens). TTFT denotes time-to-first-token. TDA decode is omitted due to the absence of an upstream cached decode kernel.

K Stress and Environment Diagnostics

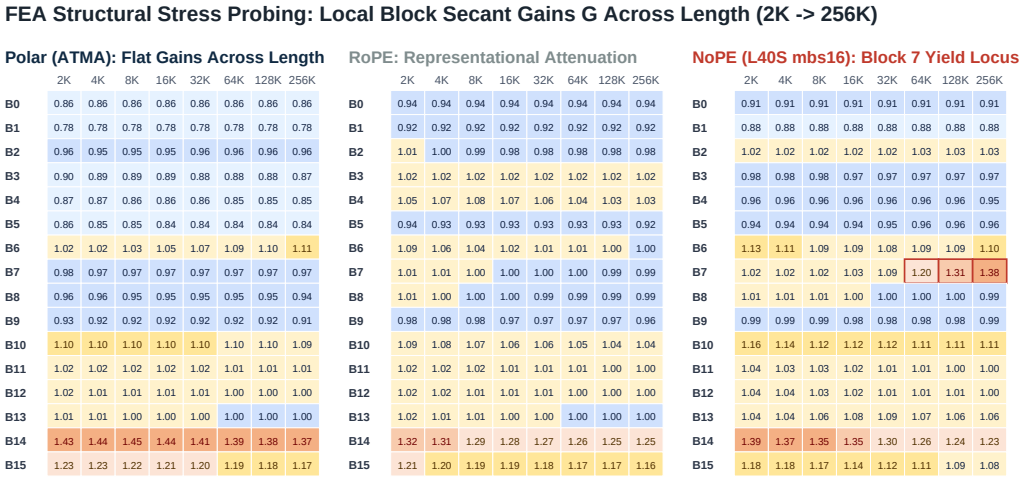


Figure 7: Block-wise sampled secant gains for the three matched attention checkpoints over 64 nested FinePDFs documents and 64 perturbation directions per document. Raven Native and Atma-Raven-Titans were not run under this checkpoint-specific probe. The diagnostic localizes changes but does not estimate a global Lipschitz constant.

Figure 7 reports the sampled gains. For each block f_i, the diagnostic samples

$$G_i = \frac{\|f_i(x+\delta x) - f_i(x)\|_{\mathrm{rms}}}{\|\delta x\|_{\mathrm{rms}}}, \qquad \|\delta x\|_{\mathrm{rms}} = 0.02\|x\|_{\mathrm{rms}}. \tag{12}$$

The maximum sampled gain is a lower bound on the local operator norm along sampled directions. It is not a tight bound on the block Lipschitz constant. In the observed NoPE collapse, changes first concentrate near blocks 6–7 and later propagate through the residual stream. Polar's sampled gains and activation moments remain comparatively flat through 256K.

L Checkpoint-Variability Diagnostic Journey

The variability finding emerged sequentially. Stage I used L4 GPUs, followed by the first 10B NoPE control on an L4. The difference became visible after primary training moved to L40S. Short-context convergence remained close (2.8126 vs 2.8160 nats; approximately 0.12% difference), while 256K NLL differed (4.07 vs 10.77 nats). Table 34 records the seven-step diagnostic sequence and the retention-horizon intervention.

Step	Observation or intervention	Diagnostic implication
1. L4 baseline	Stage I and first 10B NoPE control on L4 (mbs4); 256K NLL was 4.07.	Established baseline extrapolation, but lacked replicated device-class controls.
2. L40S transition	Promoted NoPE run on L40S (mbs16); 2K validation matched (2.8126 vs 2.8160), but 256K NLL rose to 10.77.	Showed that the short-context pretraining curves did not predict long-context behavior.
3. Evaluation replay	L4 and L40S checkpoints re-evaluated under PyTorch 2.12 and 2.13 with identical metrics.	Excluded evaluation software runtime mismatch as the source.
4. Microbatch control	NoPE retrained on L40S with mbs4; 2K validation ended at 2.8128 and 256K NLL at 9.90.	Showed that microbatch size four did not reliably prevent collapse; initialization remained uncontrolled.
5. Deterministic fixtures	Synthetic 2K stream fixtures across device classes showed zero gradient or loss divergence.	Confirmed short-context kernel execution consistency; left training trajectory unconstrained.
6. Retention parameter inspection	Evaluated learned zero-input operating point $\gamma_0 = \sigma(b_\gamma)$ across all layers and heads.	**Identified an outlier:** L40S NoPE developed an isolated Block 2 head with $b_\gamma = +13.33$ ($H_{1/2} = 21.01$M tokens), whereas L4 NoPE had max $H_{1/2} = 185.3$ tokens.
7. Single-head causal intervention	Installed a reversible inference cap ($H_{\max} = 256$ tok) on the single outlier head and re-evaluated 2K–256K.	**Tested the mechanism:** Capping the outlier head reduced 256K NLL from 12.00 to 1.14 nats and raised 256K BABILong to 39%, supporting excessive retention as a mediator of the collapse.

Table 34: Diagnostic sequence from the initial L4/L40S divergence to the retention-horizon intervention. Steps 1–5 tested evaluation artifacts and showed that reducing microbatch size did not reliably prevent the failure. Steps 6–7 identified the isolated retention-gate outlier and tested its effect.

This sequence localizes the observed checkpoint failure to one unusually long zero-input retention horizon, and the inference cap reverses much of the measured degradation. It does not establish what training event produced the outlier or whether hardware-level differences contributed to it. All five headline Stage II models were trained in the same L40S environment, providing an internally controlled comparison for those runs.