WHEN DOES A ROPE EXTRAPOLATION REPAIR HELP?

Anonymous authors
Paper under double-blind review

ABSTRACT

Context-extension repairs for rotary position embeddings (RoPE) can rescue a model whose predictive performance collapses on long documents, yet harm one that already handles long inputs, so practitioners must know in advance which case applies. We study a diagnostic that predicts when one low-budget repair helps. The repair retains RoPE, adds to each attention head a distance penalty scaled to how far back that head attends, and trains for 30 updates. The diagnostic is the matched no-penalty continuation's ratio of perplexity after a document's first 2,048 tokens (late-token perplexity) to perplexity on those initial tokens. At a ratio of at least 2, a fixed rule fires and predicts at least 5% lower late-token perplexity than that continuation; otherwise, it abstains. On 30 checkpoint–length cells from ten public checkpoints, including five named before evaluation, the rule fires on 20 and all 20 improve; it abstains on exactly the 10 cells without a reduction, including native long-context Qwen3 checkpoints, where the repair raises perplexity. On three independently trained 124M parents, the repair lowers 16K late-token perplexity from 168–182 to about 25 in 74 GPU-seconds, at a cost of about 0.1 in in-window (PG19) perplexity; measured slopes beat three calibration-free schedules on late-token perplexity for every parent. In a key-swap control, a correct key just before the question raises the repaired models' answer log-likelihood 7.3–8.2 nats more than a wrong key. On Pythia-1b, the repair has the lowest in-window perplexity of the compared extensions. A single paired run thus makes the decision to apply this repair testable.

1 INTRODUCTION

Language models are often applied to documents longer than those they were trained on. For models with rotary position embeddings (RoPE) (Su et al., 2021), which encode token order by rotating query and key vectors, the usual response is a positional repair followed by brief adaptation. Such a repair can keep perplexity from exploding beyond the training length (Chen et al., 2023; Peng et al., 2023). On a checkpoint that already handles long inputs, a repair can instead raise perplexity, as on two native long-context Qwen3 checkpoints (Section 4.2). For a practitioner with a fixed checkpoint and a small adaptation budget, the first question is therefore whether a given repair will help.

Extension methods are usually evaluated on checkpoints that degrade beyond their training length (Chen et al., 2023; Peng et al., 2023; Jin et al., 2024). Such evaluations show that a method can help a model that needs it, but they provide no criterion measured on a new checkpoint for predicting whether a specified repair will lower long-document loss or should be withheld.

We study this question for one additive repair of RoPE (Figure 1a). The repair keeps the trained rotations and adds an ALiBi-style penalty (Press et al., 2021), i.e., a bias on each attention logit that decreases linearly with query–key distance. Its key design choice is a per-head slope calibrated from the head's measured *attention horizon*: the 90th percentile of the distances over which the head spreads its attention mass. We call the result the *horizon repair*. It requires one forward measurement pass, 30 continuation updates, and a fixed bias at inference, served through a fused kernel's ALiBi interface. We evaluate late-token perplexity and content-paired retrieval.

We predict when this repair helps from one measured ratio (Figure 1b). We split each long document after its first 2,048 tokens; perplexity on those tokens is *initial-token perplexity*, and on the remaining tokens *late-token perplexity*. A *matched continuation* control receives the same initialization, token stream, and 30 updates as the repair, without the penalty. The diagnostic B is the control's late-

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.

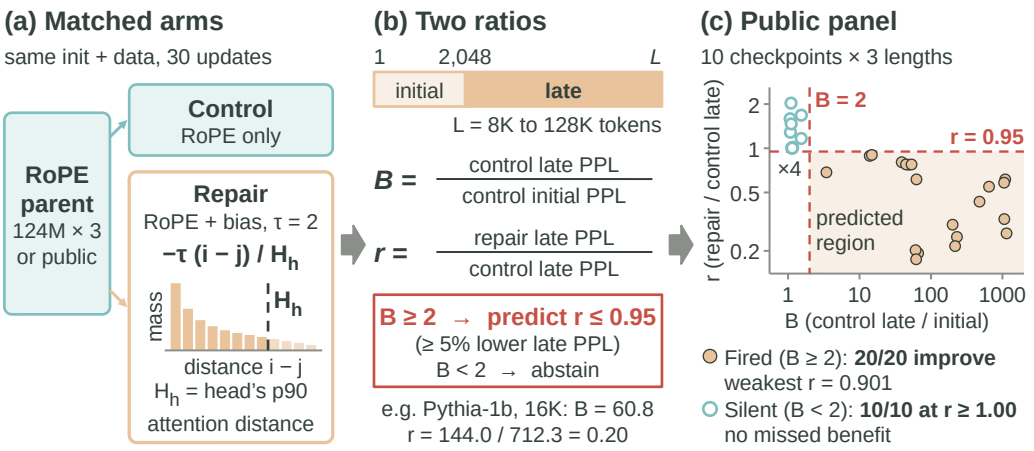


Figure 1: Overview of the repair, the diagnostic, and the public-panel result. (a) A RoPE parent (a 124M PG19 development parent or a public checkpoint) is continued for 30 updates in two arms with shared initialization and data: a control, and a repair that adds the per-head attention-logit bias $-\tau(i-j)/H_h$ ($\tau = 2$; H_h is the 90th percentile of head h's attention distance). (b) Each length-L document is split after token 2,048. The diagnostic B is the control's late/initial perplexity (PPL) ratio, the outcome r the repair/control late-token PPL ratio, and the fixed rule predicts $r \leq 0.95$ when $B \geq 2$ and abstains otherwise (e.g., Pythia-1b, 16K: $B = 60.8$, $r = 144.0/712.3 = 0.20$). (c) On 30 public checkpoint–length cells (10 checkpoints $\times$ 3 lengths; five named before evaluation: three fired, two silent), all 20 fired cells improve (weakest $r = 0.901$) and all 10 silent cells (Qwen3, Phi-3, and SmolLM2 at 8K; "$\times 4$" marks four coincident cells at $r \approx 1.00$) have $r \geq 1.00$, so the rule separates every cell. Footer: on the three development parents, 30 updates lower 16K late-token perplexity from 168–182 to about 25 in 74 GPU-seconds, at about 0.1 in-window (PG19) perplexity.

token perplexity divided by its initial-token perplexity at the evaluated length, and the outcome r is the repair's late-token perplexity divided by the control's (Section 2.1). A ratio of 2 or more marks a checkpoint whose perplexity deteriorates after the first 2,048 tokens and so needs a long-context fix. A fixed rule turns this need into a falsifiable prediction: whenever B is at least 2, the repair lowers late-token perplexity by at least 5% relative to the matched control ($r \leq 0.95$). We say that the rule *fires* on such a checkpoint–length cell and is *silent* otherwise.

Figure 1c summarizes the central result. On ten public checkpoints at three lengths each, five of the checkpoints named before evaluation, the rule fires on 20 of the 30 checkpoint–length cells and all 20 improve; it is silent on the other 10, none of which improves, including two native-32K Qwen3 checkpoints on which the repair raises late-token perplexity by up to 2.03$\times$ (Section 4.2). On three independently trained 124M parents, the repair lowers 16K late-token perplexity from 168–182 to about 25 in 74 GPU-seconds (Section 4.1). On Pythia-1b, SelfExtend (Jin et al., 2024), which regroups distant positions without training, and YaRN (Peng et al., 2023), which rescales rotary frequencies and is continued for the same 30 updates, also lower 16K late-token perplexity (to 27.1 and 17.3; the horizon repair reaches 144.0), consistent with B measuring need. The horizon repair keeps the lowest in-window perplexity, so the choice among methods follows the target metric (Section 4.3).

Our contributions are as follows.

- **A diagnostic with a fixed decision rule, tested on a public panel.** We read a late-to-initial perplexity ratio from a matched continuation, fix a rule that predicts when the repair lowers late-token perplexity, and test it on 30 cells from ten public checkpoints, where it separates every cell (weakest fired ratio 0.90; Sections 2.1 and 4.2).

- **A horizon-calibrated repair.** We set each head's penalty slope from its measured attention horizon and keep RoPE; the benefit replicates on three parents at about 0.1 points of in-window (PG19) cost, with lower late-token perplexity than three calibration-free schedules (Sections 4.1 and 5).
- **A cost profile against standard extensions.** We compare the repair with SelfExtend and YaRN: on Pythia-1b it has the lowest in-window cost (PG19 perplexity 14.75, versus 15.32 for continued YaRN and 16.28 for SelfExtend), and on the development model it decodes 131.0 tokens/s at 32K, versus 117.9 for SelfExtend on the same reference path (Sections 4.1 and 4.3).
- **Content-paired retrieval controls.** We separate content-selective from non-selective retrieval gains by replacing the key's digits while keeping complete 16,380-token scenes fixed; when the key directly precedes the question, the repair's content selectivity is $+7.76$ nats (Section 5).

2 Diagnostic and horizon-calibrated repair

2.1 A directional rule for a specified intervention

For each long document, split prediction targets after the first 2,048 tokens. Let $P_{\text{initial}}(L)$ and $P_{\text{late}}(L)$ be the exponentiated mean cross-entropies in these two regions. For the matched arms of Figure 1(a–b), the diagnostic B and the outcome r are

$$B(L) = \frac{P_{\text{late}}^{\text{control}}(L)}{P_{\text{initial}}^{\text{control}}(L)}, \qquad r(L) = \frac{P_{\text{late}}^{\text{repair}}(L)}{P_{\text{late}}^{\text{control}}(L)}. \tag{1}$$

The tested prediction is $B(L) \geq 2 \Rightarrow r(L) \leq 0.95$ for the 30-update, $\tau = 2$ repair. The rule *fires* on a checkpoint–length cell when $B(L) \geq 2$ and is *silent* otherwise. The cutoff was fixed before the outcomes of the prospective checkpoints were measured; earlier and development checkpoints are reported separately (Section 4.2). When the rule is silent, it abstains.

The 2,048-token split is fixed across the public panel, including models with longer native context windows, so B measures long-document deterioration under one common evaluation protocol.

2.2 Measuring the slopes

For head h, pool its post-softmax attention mass by query–key distance and take the 90th percentile H_h. The development family uses 48 held-out 2,048-token PG19 windows; public-model maps use 24. Each query row, sampled at stride four from position 32, contributes unit mass, so H_h is a quantile of a pooled distribution rather than an average of row quantiles. A zero horizon is imputed by its layer's nonzero median; such entries must comprise at most 10% of the map (Appendix C).

The attention-logit intervention is

$$b_h(i, j) = -\frac{\tau}{H_h}(i - j), \qquad j < i. \tag{2}$$

The penalty scale τ is shared by all heads, and RoPE frequencies remain unchanged. There are no additional learned parameters; the measured slopes persist at inference through the fused kernel's ALiBi interface (Dao et al., 2022; Dao, 2023), with no input-adaptive computation at deployment. Thirty continuation updates are 1% of the parent family's training updates. A separate 300-update study tests whether more adaptation reduces cross-domain cost; its $\tau = 1$ point was selected from the measured dose ladder (Appendix C).

2.3 What the intervention does

The additive term supplies a content-independent recency preference. It can suppress inappropriate distant attention without changing the trained phase geometry. Replacing each head's rotary base by $H_h/(2\pi)$ instead causes large phase changes in our development model, so we retain RoPE.

An idealized Gaussian-content calculation makes this precise at score level: it separates the oscillatory score covariance of the rotary channel from the deterministic distance penalty, whose strength does not depend on content (Appendix H).

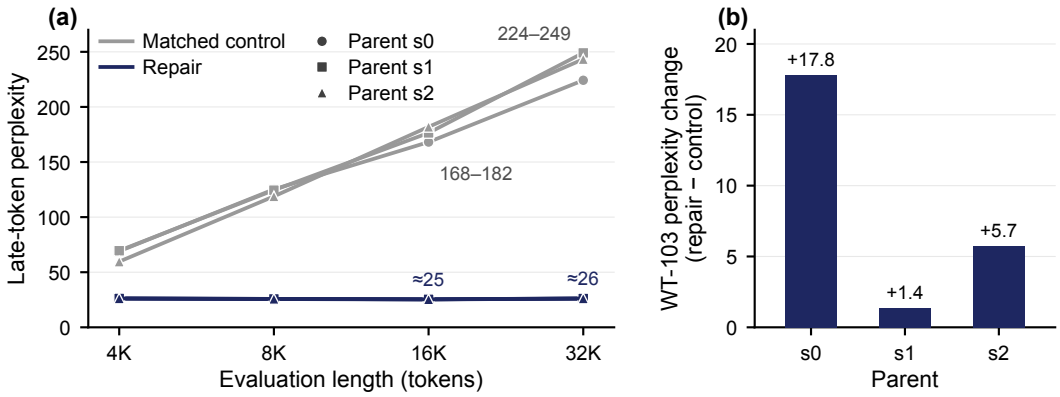


Figure 2: Matched 30-update outcomes in three independently trained parents (s0–s2, training seeds 0–2). (a) Late-token perplexity against evaluation length for the matched control (grey) and the repair (navy); markers identify the parent. Control perplexity grows from 60–70 at 4K to 168–182 at 16K and 224–249 at 32K, while the repair stays at about 25–26 at every length (its three lines coincide), so the late-token benefit replicates in every parent. (b) WikiText-103 perplexity of each repaired model minus that of its matched control. Each line or bar is a single training run. The 32K values come from a separate 32K evaluation of the same checkpoints.

3 Experimental design

The development family consists of three independently trained 124M RoPE transformers, using a decoder-only autoregressive architecture (Radford et al., 2019), each trained on 786M PG19 tokens at context 2,048 (Rae et al., 2019). Within each parent, repair and no-repair continuation share the initialization, token stream, update count, and remaining optimizer schedule. Their differences therefore isolate the combined slope intervention and its training response. The same slope assignment, trained from scratch, provides an additional reference point with its own training budget.

We measure *in-window perplexity* on held-out 2,048-token windows, in-domain on PG19 validation text and zero-shot on WikiText-103 (WT-103), and late-token perplexity at longer windows; an increase in in-window perplexity is an *in-window cost*. Perplexity (PPL) is the exponential of mean token cross-entropy over the stated region, not mean per-window perplexity. The *needle margin* is the answer's summed negative log-likelihood given the question alone minus that given the full retrieval scene (nats; higher is better); in the fixed-grid retrieval battery (scenes of about 9.3K tokens), requested depth 0.9 places the key directly before the question (Appendix G). The family long-document evaluation uses up to 24 PG19 test books; each public-model panel uses up to eight sufficiently long books. The public panel covers ten OLMo-2, Pythia, TinyLlama, SmolLM2, Phi-3, and Qwen3 checkpoints (OLMo et al., 2024; Biderman et al., 2023; Zhang et al., 2024; Allal et al., 2025; Abdin et al., 2024; Yang et al., 2025), listed in Table 1. They have different native training contexts and tokenizers. We compare methods within a checkpoint and do not pool their perplexities. The compared alternatives are YaRN (Peng et al., 2023), which rescales rotary frequencies and is evaluated both without training and after the same 30-update continuation, and SelfExtend (Jin et al., 2024), which maps distant positions onto grouped positions without training.

Retrieval intervals are pointwise paired t intervals over 16 contents at fixed weights. Appendix G details the protocols and realized scene lengths.

4 When the repair helps

4.1 Replicated loss reduction across three parents

Does the repair reliably reduce late-token loss across independently trained parents, and at what in-window cost? The repair reduces late-token loss in all three parents (Figure 2). At 16K, controls score $168.1/176.4/181.9$, while repaired models score $25.3/25.4/25.4$. The corresponding in-

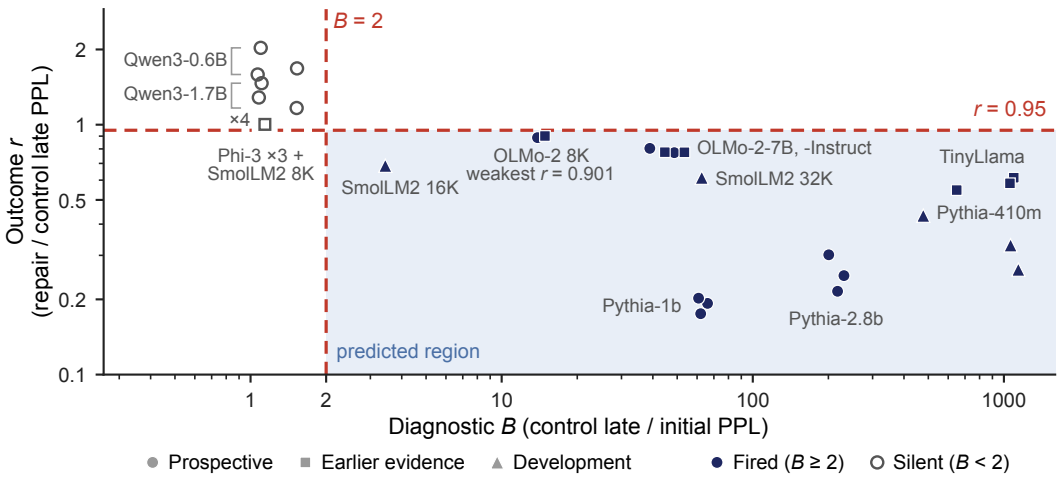


Figure 3: The measured public-checkpoint panel (values in Table 1). Each point is one checkpoint–length cell. Both axes are logarithmic. Marker shape identifies the evidence cohort: circles: *Prospective* (OLMo-2-7B-Instruct, Pythia-1b, Pythia-2.8b, Qwen3-0.6B, Qwen3-1.7B; named before evaluation), squares: *Earlier evidence* (OLMo-2-7B, TinyLlama-1.1B, Phi-3-mini-4k; measured under an earlier scoring rule, Appendix J), and triangles: *Development* (Pythia-410m, SmolLM2-1.7B; used while developing the rule). Filled markers are fired cells ($B \geq 2$), open markers silent cells ($B < 2$). Red dashed lines mark the fixed diagnostic threshold $B = 2$ and the outcome threshold $r = 0.95$; the shaded area is the predicted region. Brackets span each Qwen3 model's cells; the OLMo-2 8K label gives the weakest fired outcome ($r = 0.901$). The four silent cells of Phi-3 (8K/16K/32K) and SmolLM2 at 8K coincide at $B \approx 1.1$, $r \approx 1.00$ (marked $\times 4$). All 30 cells are shown; each is a single paired run.

window (PG19) costs are $+0.12/+0.11/+0.12$. At 32K, repair perplexities remain $26.0/26.1/26.2$, versus 224/249/243 for the matched controls, so the benefit persists at twice the 16K length.

On WikiText-103, the 30-update repair changes perplexity by $+17.8/+1.4/+5.7$ points relative to its matched control. Longer adaptation lowers this cost: in the 300-update study, $\tau = 1$ reduces the seed-0 matched cost to $+0.30$ while retaining low late-token perplexity (Appendix C, Figure 4).

The 30-update seed-0 continuation takes 74 GPU-seconds, excluding the separate horizon-measurement sweep; the 300-update setting takes 646 GPU-seconds. On the reference inference path, the repaired model decodes 131.0 tokens/s at 32K, versus 117.9 for SelfExtend; both retain their positional computations at inference. Served through FlashAttention's native ALiBi-slope interface, the repaired model decodes $206.5/200.0/207.5$ tokens/s at 4K/16K/32K, and its end-to-end deviation from the reference path is consistent with the measured kernel noise floor (Appendix E).

4.2 The diagnostic separates all 30 public cells

Does the fixed rule separate the public checkpoints that benefit from those that do not? Figure 3 and Table 1 show all 30 public checkpoint–length cells. The rule fires on 20 cells; every fired cell satisfies $r \leq 0.95$ (largest $r = 0.901$), and every silent cell has r between 1.00 and 2.03. Five checkpoints were named before their outcomes were measured. On three of them, OLMo-2-7B-Instruct, Pythia-1b, and Pythia-2.8b, the rule fires at every length and every cell satisfies the predicted direction; on Qwen3-0.6B and Qwen3-1.7B it is silent at every length. The other five checkpoints, measured under an earlier scoring rule or used during development, follow the same pattern and are listed as separate cohorts (Appendix J); among them, the development checkpoint Pythia-410m improves at all three lengths ($r = 0.43/0.33/0.26$).

Both native-32K Qwen3 checkpoints stay below the diagnostic cutoff at 32K, 64K, and 128K (B between 1.07 and 1.53). Applying the repair anyway raises late-token perplexity by factors of 1.59/2.03/1.68 at 0.6B and 1.28/1.47/1.17 at 1.7B. Phi-3 is also silent, with $r = 1.00$ at all three

Table 1: Full public-checkpoint panel (eight PG19 books per cell). B is the diagnostic's range over the three evaluated lengths (three significant figures below 1,000, integers above); r is the repair/control late-token perplexity ratio at each length (three decimals; same paired result files as Figure 3). Parenthesized r values are silent cells ($B < 2$) on which the repair was applied anyway. "Weakest $1 - r$" is the smallest relative reduction among a row's fired cells; the rule predicts at least 0.05, and every fired row meets it. Cohorts are defined in Figure 3.

			r (repair / control)			
Cohort	Checkpoint	B range	8K	16K	32K	Weakest $1 - r$
Prospective	OLMo-2-7B-Instruct	13.9–48.7	0.888	0.804	0.773	0.112
	Pythia-1b	60.8–66.0	0.193	0.202	0.175	0.798
	Pythia-2.8b	201–231	0.302	0.249	0.215	0.698
	Qwen3-0.6B-Base†	1.07–1.53	(1.588)	(2.030)	(1.680)	—
	Qwen3-1.7B-Base†	1.08–1.53	(1.285)	(1.466)	(1.166)	—
Earlier evidence	OLMo-2-7B	14.9–53.4	0.901	0.776	0.775	0.099
	TinyLlama-1.1B	649–1095	0.614	0.583	0.547	0.386
	Phi-3-mini-4k	1.12–1.15	(1.003)	(1.003)	(1.003)	—
Development	Pythia-410m	477–1141	0.433	0.329	0.263	0.567
	SmolLM2-1.7B	1.15–62.6	(1.015)	0.685	0.614	0.315

†Qwen3-0.6B/1.7B in the text; lengths are 32K, 64K, and 128K instead of 8K, 16K, and 32K.

Table 2: Horizon repair, SelfExtend, and YaRN on the prospective checkpoint Pythia-1b (perplexity, lower is better). The first two rows are references: the parent and the matched continuation, which receives the repair's 30 updates without the penalty. Continued arms share the training budget; training-free SelfExtend is a distinct cost point. Bold marks the best extension arm in each column. Full retrieval and cost measurements are in Appendix A (Tables 5 and 7).

	In-window PPL ↓		Late-token PPL ↓	
Pythia-1b arm	PG19	WT-103	16K	32K
Parent	14.233	12.370	758.2	777.9
Matched continuation	14.648	12.731	712.3	725.9
Horizon repair	**14.746**	**13.059**	144.0	127.2
SelfExtend, no adaptation	16.275	13.228	27.1	125.5
YaRN, evaluation only	56.198	61.901	238.2	500.2
YaRN, 30-update continuation	15.317	14.206	**17.3**	**93.8**

lengths and small positive in-window costs. These cells show the value of abstention: the rule withholds the repair from checkpoints that already handle long inputs.

Borderline cells. Only SmolLM2-1.7B approaches the cutoff from both sides: at 8K it is silent ($B = 1.15$, $r = 1.015$), whereas at 16K it fires with $B = 3.44$ and improves ($r = 0.685$). The rule matches the outcome at both lengths. Every other fired cell has $B > 13$ and every other silent cell $B < 1.54$ (Table 1), so any cutoff between 1.54 and 3.44 classifies these 30 cells identically. The panel thus spans two well-separated regimes, checkpoints whose predictive performance collapses at the evaluated length and native-long-context checkpoints whose perplexity stays nearly flat, and the rule assigns each cell to its regime.

4.3 Comparison with SelfExtend and YaRN

How does the repair compare with standard extensions on a checkpoint where the rule fires? On Pythia-1b, the repair reduces 16K late-token perplexity from 712.3 to 144.0; SelfExtend reaches 27.1 without adaptation and continued YaRN reaches 17.3 (Table 2). The horizon repair has the smallest in-window cost among these extensions: PG19 validation perplexity is 14.746 against 14.648 for the matched continuation, versus 16.275 for SelfExtend and 15.317 for continued YaRN, and WT-103 perplexity is 13.059 versus 13.228 and 14.206. On the development model it also de-

Table 3: Advantage of the measured slope map over three calibration-free schedules at the same dose (30 updates, $\tau = 2$) on each parent; positive values favor the measured map. Late-token PPL at 16K: schedule minus measured map. Needle margin (nats; fixed-grid battery, requested depth 0.9, where the key directly precedes the question in 9.3K-token scenes; Appendix G): measured map minus schedule. WT-103 cost: the schedule's excess over the matched control minus the measured map's (negative for the uniform schedule, whose in-window cost is lower). Each cell is one run.

		Advantage of the measured map (+ = measured better)		
Schedule	Parent	Late-token PPL, 16K	Needle (nats)	WT-103 cost (PPL)
Geometric ladder	s0	$+21.1$	$+11.5$	$+23.8$
	s1	$+8.1$	$+1.7$	$+44.8$
	s2	$+9.8$	$+1.4$	$+14.9$
ALiBi slopes	s0	$+5.8$	$+5.0$	$+816.9$
	s1	$+6.3$	$+3.4$	$+820.7$
	s2	$+4.9$	$+1.3$	$+370.4$
Uniform horizon	s0	$+38.4$	$+16.5$	-18.1
	s1	$+44.3$	$+18.5$	-5.4
	s2	$+34.6$	$+10.4$	-4.6

codes about 11% faster at 32K than SelfExtend on the reference path (131.0 versus 117.9 tokens/s; Section 4.1). The repair thus occupies the low-in-window-cost end of the trade-off, and the choice among methods follows the target metric and cost budget (Jin et al., 2024; Peng et al., 2023).

The same comparison supports reading B as a need signal. On Pythia-1b, whose cells all fire, SelfExtend and both YaRN arms also lower 16K late-token perplexity far below the parent's 758.2 (to 27.1, 238.2, and 17.3). The fixed rule adds a falsifiable prediction for one repair.

On the development model, the horizon repair has the lowest 16K late-token perplexity among the 30-update and training-free arms (25.29 versus 25.78 for continued YaRN, read under its own operator, and 27.98 for SelfExtend) and about half of continued YaRN's at 32K (26.04 versus 52.86; Table 6, which also lists a plain-head reading of the YaRN checkpoint).

5 What horizon calibration contributes

Measured slopes outperform calibration-free schedules at long range. A geometric horizon ladder from 64 to 16,384 tokens, the canonical ALiBi slope set, and a uniform horizon of 2,048 all underperform the measured map at the same 30-update dose on each parent (Table 3). The measured map's 16K late-token perplexity advantage ranges from $+4.9$ to $+44.3$ points, and the needle margin also favors the measured map in all nine cells. The schedules fall short in different ways: canonical ALiBi disturbs in-window statistics so strongly that 30 updates cannot absorb it (WT-103 perplexity $+370$ to $+821$ points above the measured map's), whereas the uniform schedule is benign in-window but worst at long range. Against these schedules, the measured slopes give the lowest late-token perplexity and the largest needle margin at the 30-update dose on every parent (details in Appendix B).

Aggregate map structure, not head identity. Sorted horizon profiles are far more stable than head identities. Across the three independently trained parents, every pair of sorted profiles correlates at ≥ 0.998 and coverage stays within $[0.708, 0.771]$, whereas per-head horizon ranks correlate only at Spearman 0.25–0.31. The aggregate profile also survives a ten-fold change of training budget: on a parent trained on 78.6M instead of 786M tokens, the sorted profile correlates at 0.986–0.989 with the full-budget profiles and coverage is 0.771 (Appendix C). Shuffling a map while preserving its value multiset retains much of the long-window benefit, at a larger in-window cost. These findings favor an aggregate interpretation, in which two aggregate ingredients characterize a map: its value distribution and its *coverage* C, the share of nonzero-horizon heads whose horizon lies between 300 and 1,200 tokens (Appendices B and C).

Table 4: Replication across the three development parents and zero-diagnostic map transfer (30 updates, $\tau = 2$). Costs are repair minus matched control (perplexity points); the needle margin (nats, higher is better) is read at requested depth 0.9 of the fixed-grid battery, where the key directly precedes the question in 9.3K-token scenes (Appendix G). The last two rows apply the seed-0 map to the siblings without measuring their horizons, each against its own matched control. Each row is one trained run.

		Cost (Δ PPL)		Late-token PPL, 16K		Needle
Parent	Slope map	PG19 val.	WT-103	Control	Repair	(nats)
s0	own	$+0.12$	$+17.8$	168.1	25.3	$+15.4$
s1	own	$+0.11$	$+1.4$	176.4	25.4	$+17.3$
s2	own	$+0.12$	$+5.7$	181.9	25.4	$+17.5$
s1	s0, borrowed	$+0.13$	$+1.8$	176.4	25.5	$+14.1$
s2	s0, borrowed	$+0.18$	$+6.1$	181.9	25.8	$+17.4$

Replication and zero-diagnostic map transfer. Table 4 repeats the headline configuration on all three parents and applies the seed-0 map, unchanged, to the two siblings without measuring their horizons. Each own map lowers 16K late-token perplexity from 168–182 to about 25 and raises the needle margin to $+15.4/+17.3/+17.5$ nats. The borrowed map does nearly as well: late-token perplexity 25.5 and 25.8, needle margins of $+14.1$ and $+17.4$ nats, and PG19 in-window costs of $+0.13$ and $+0.18$. One sweep thus serves sibling checkpoints in the family (Appendix D).

Coverage marks the development-scale regime. The development maps have C between 0.708 and 0.771, whereas the five public maps with at least 1B parameters whose coverage was probed (OLMo-2-7B, Qwen3-4B/8B, Phi-3-mini, SmolLM2-1.7B) all have $C \leq 0.087$, and Pythia-410m has $C = 0.161$. We use this separation as a *coverage gate*: the full development-scale outcome applies to maps with family-like coverage. The directional rule is evaluated independently of the gate and holds on all 20 fired public cells. On the low-coverage Qwen3-4B-Base, a 30-update repair with its penalty magnitude capped at 0.25 left all nine retrieval and tracking cells of a held-out decode battery unchanged (Appendix F).

Content-paired retrieval measurement. The dedicated distance experiment uses complete 16,380-token scenes (Appendix G). In the dedicated experiment, a true-key versus swapped-key comparison holds scene structure fixed and resolves the repair's advantage over the matched control into a content-selective and a non-selective part. With no filler between key and question, the repair's content selectivity is $+7.76$ nats on seed 0 ($+7.29$ and $+8.18$ on the other two parents), and on seed 0 it remains $+5.38$ nats at 1,051 filler tokens. The decomposition covers every filler distance up to 8,192 tokens (Appendix I); decoded-answer results for the public comparison are in Table 7.

6 Related work

Positional geometry and locality. Attention-based language models need a representation of sequence order (Vaswani et al., 2017; Su et al., 2021). ALiBi, KERPLE, and FIRE impose distance-dependent score structure; XPos modifies rotary behavior to support extrapolation (Press et al., 2021; Chi et al., 2022a; Li et al., 2023; Sun et al., 2022). Adaptive spans and content-dependent position or forgetting mechanisms instead vary locality with the model or input (Sukhbaatar et al., 2019; Golovneva et al., 2024; Lin et al., 2025). Alternative position algebras (group-representational encodings; Fourier–jet–affine position spaces) and attention-form variants (the Differential Transformer) chart the neighboring design space (Zhang et al., 2025; Zhang, 2026; Ye et al., 2024). Positional schemes have been argued to set the default spectral algebra of attention (Li, 2026). Our operator is an additive linear bias from this existing family; the contribution lies in its measured per-head calibration and in a conditional treatment prediction.

Extending a trained model. Position interpolation, NTK-aware community rescalings, and YaRN adapt rotary coordinates; dilated attention scales the per-head pattern itself; LongRoPE and LongRoPE2 optimize rescaling and continuation choices (Chen et al., 2023; bloc97, 2023; Peng et al.,

2023; Ding et al., 2023; 2024; Shang et al., 2025). SelfExtend provides a training-free grouped-position alternative, while DPE manipulates selected positional dimensions and AdaRoPE studies head-specific frequencies and scaling (Jin et al., 2024; Lu et al., 2025; Wang et al., 2026). DoPE identifies problematic heads through a different spectral statistic (Xiong et al., 2025). Our measured horizon is a distance quantile, not a norm, entropy, or frequency-band estimate; we retain rotations and add a recency penalty. We compare against SelfExtend and YaRN at matched budgets.

Interpreting extrapolation. Receptive-field analysis connects positional biases with accessible context (Chi et al., 2022b), and mechanistic analyses ask whether rotary scoring itself prevents or degrades retrieval heads (Bayram, 2026). Work on positional generalization and RoPE's learned frequency use shows that useful behavior cannot be inferred from distance decay alone (Kazemnejad et al., 2023; Barbero et al., 2024). Frequency Bands in RoPE further relates the dominant norm band to the base and training length and reports an interpolation–extrapolation trade-off (Oka et al., 2026). These studies motivate our head-horizon controls and fixed-checkpoint perturbations.

Long-context evaluation. StreamingLLM preserves stable generation with a bounded attention context; Landmark Attention offers a different route to long-range access (Xiao et al., 2023; Mohtashami & Jaggi, 2023). Lost in the Middle, retrieval-head analyses, RULER, and LongBench evaluate long-context information use (Liu et al., 2023; Wu et al., 2024; Hsieh et al., 2024; Bai et al., 2023). We use swapped-key and decoded-answer controls.

7 Discussion

Separating need, treatment effect, and task benefit makes the decision to apply a low-budget repair testable: the late-token ratio identifies the cells where the specified repair helps, and the rule abstains exactly on the evaluated cells where the repair brings no reduction. Horizon calibration makes the repair effective on three independently trained parents, ahead of calibration-free schedules at long range and with the lowest in-window perplexity among the extensions compared on Pythia-1b; content-paired controls quantify its retrieval effect.

Limitations and future work. The public panel contains 30 checkpoint–length cells from ten checkpoints, with shared weights and documents across lengths; broader prospective tests are needed, and the full recipe is demonstrated on one development family. The diagnostic B requires a trained matched continuation, so it is not a training-free predictor, and the rule addresses one specified repair. Lower loss does not establish decoded retrieval: the content-paired evidence is teacher-forced, the remaining gain is non-selective from 4,096 filler tokens on, and decoded retrieval on Pythia-1b favors SelfExtend and YaRN (Table 7). At 4B, the capped repair lowers accuracy on one 32K reasoning cell across four item draws (Appendix F). Future work should test cheaper diagnostics, selection among repairs, content-addressable access, and the emergence of head horizons (Appendix K).

Ethics Statement

This work trains and evaluates language models on publicly available corpora, benchmarks, and model checkpoints. It involves no human-subject experiments. Public corpora can contain personal or sensitive content, and the work shares the data-related and model-use risks common to language-model research.

Reproducibility Statement

The supplementary material (a zip archive) provides code, configurations, launch scripts, analysis scripts, and result files, together with a README linking reported tables and figures to their producing scripts. Experimental settings are specified in the main text and the appendix.

AI Use Statement

During the project, vibe coding tools (Codex and Claude Code) were used to help develop the code for the methods and experiments. All such work was conducted under human supervision and underwent human review. During manuscript preparation, AI tools were used to polish and revise the text and figures.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, 2024. URL https://arxiv.org/abs/2404.14219.

Habibullah Akbar. ATMA: Long-Context Language Modeling via Polar Attention and Gated-Delta Compression Memory, 2026. URL https://arxiv.org/abs/2606.25156.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, et al. SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model, 2025. URL https://arxiv.org/abs/2502.02737.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, et al. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding, 2023. URL https://arxiv.org/abs/2308.14508.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and Round We Go! What makes Rotary Positional Encodings useful?, 2024. URL https://arxiv.org/abs/2410.06205.

Cengizhan Bayram. Does RoPE Prevent or Degrade Retrieval Heads? A Mechanistic Analysis Across Model Families, 2026. URL https://arxiv.org/abs/2606.21249.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, et al. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling, 2023. URL https://arxiv.org/abs/2304.01373.

bloc97. NTK-aware scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation, 2023. Reddit post, r/LocalLLaMA, July 2023. Accessed 2026-08-30.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending Context Window of Large Language Models via Positional Interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Ta-Chung Chi, Ting-Han Fan, Peter J. Ramadge, and Alexander I. Rudnicky. KERPLE: Kernelized Relative Positional Embedding for Length Extrapolation, 2022a. URL https://arxiv.org/abs/2205.09921.

Ta-Chung Chi, Ting-Han Fan, Alexander I. Rudnicky, and Peter J. Ramadge. Dissecting Transformer Length Extrapolation via the Lens of Receptive Field Analysis, 2022b. URL https://arxiv.org/abs/2212.10356.

Tri Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, 2022. URL https://arxiv.org/abs/2205.14135.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. LongNet: Scaling Transformers to 1,000,000,000 Tokens, 2023. URL https://arxiv.org/abs/2307.02486.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens, 2024. URL https://arxiv.org/abs/2402.13753.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual Position Encoding: Learning to Count What's Important, 2024. URL https://arxiv.org/abs/2405.18719.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models?, 2024. URL https://arxiv.org/abs/2404.06654.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. LLM Maybe LongLM: Self-Extend LLM Context Window Without Tuning, 2024. URL https://arxiv.org/abs/2401.01325.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The Impact of Positional Encoding on Length Generalization in Transformers, 2023. URL https://arxiv.org/abs/2305.19466.

Hengyu Li. Fingerprint, Not Blueprint: How Positional Schemes Set the Default Spectral Algebra of Attention, 2026. URL https://arxiv.org/abs/2607.06621.

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, et al. Functional Interpolation for Relative Positions Improves Long Context Transformers, 2023. URL https://arxiv.org/abs/2310.04418.

Zhixuan Lin, Evgenii Nikishin, Xu Owen He, and Aaron Courville. Forgetting Transformer: Softmax Attention with a Forget Gate, 2025. URL https://arxiv.org/abs/2503.02130.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts, 2023. URL https://arxiv.org/abs/2307.03172.

Yi Lu, Wanxu Zhao, Xin Zhou, Chenxin An, Chenglong Wang, Shuo Li, Yuming Yang, Jun Zhao, et al. Effective Length Extrapolation via Dimension-Wise Positional Embeddings Manipulation, 2025. URL https://arxiv.org/abs/2504.18857.

Amirkeivan Mohtashami and Martin Jaggi. Landmark Attention: Random-Access Infinite Context Length for Transformers, 2023. URL https://arxiv.org/abs/2305.16300.

Yui Oka, Itsumi Saito, Kyosuke Nishida, and Kuniko Saito. Frequency Bands in RoPE: Base Frequency and Context Length Shape the Interpolation–Extrapolation Trade-off. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/993dbaec0418a6449090f1debbcb8844-Abstract-Conference.html.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, et al. 2 OLMo 2 Furious, 2024. URL https://arxiv.org/abs/2501.00656.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient Context Window Extension of Large Language Models, 2023. URL https://arxiv.org/abs/2309.00071.

Ofir Press, Noah A. Smith, and Mike Lewis. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation, 2021. URL https://arxiv.org/abs/2108.12409.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Technical Report, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. Compressive Transformers for Long-Range Sequence Modelling, 2019. URL https://arxiv.org/abs/1911.05507.

Ning Shang, Li Lyna Zhang, Siyuan Wang, Gaokai Zhang, Gilsinia Lopez, Fan Yang, Weizhu Chen, and Mao Yang. LongRoPE2: Near-Lossless LLM Context Window Scaling, 2025. URL https://arxiv.org/abs/2502.20082.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, 2021. URL https://arxiv.org/abs/2104.09864.

Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. Adaptive Attention Span in Transformers, 2019. URL https://arxiv.org/abs/1905.07799.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, et al. A Length-Extrapolatable Transformer, 2022. URL https://arxiv.org/abs/2212.10554.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017. URL https://arxiv.org/abs/1706.03762.

Shaowen Wang, Yuke Zheng, Tansheng Zhu, Shuang Chen, Shaofan Liu, Suncong Zheng, and Jian Li. AdaRoPE: Not All Attention Heads Should Rotate and Scale Equally, 2026. URL https://arxiv.org/abs/2607.19363.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval Head Mechanistically Explains Long-Context Factuality, 2024. URL https://arxiv.org/abs/2404.15574.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks, 2023. URL https://arxiv.org/abs/2309.17453.

Jing Xiong, Liyang Fan, Hui Shen, Zunhai Su, Min Yang, Lingpeng Kong, and Ngai Wong. DoPE: Denoising Rotary Position Embedding, 2025. URL https://arxiv.org/abs/2511.09146.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, et al. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential Transformer, 2024. URL https://arxiv.org/abs/2410.05258.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An Open-Source Small Language Model, 2024. URL https://arxiv.org/abs/2401.02385.

Yaobo Zhang. PJ-RoPE: A Fourier-Jet-Affine Position Space for Relative Attention, 2026. URL https://arxiv.org/abs/2606.05345.

Yifan Zhang, Zixiang Chen, Yifeng Liu, Zhen Qin, Huizhuo Yuan, Kangping Xu, Yang Yuan, Quanquan Gu, et al. Group Representational Position Encoding, 2025. URL https://arxiv.org/abs/2512.07805.

A COMPLETE NUMERICAL COMPARISONS

The following tables retain all measured outcomes from the development and public comparison panels. Different baselines, update counts, and scene realizations are stated in their captions. Cells marked — were not measured and are not imputed. Checkpoint–length cells share weights and are not independent training replications. The calibration-free schedule comparison (Table 3) and the public-checkpoint panel (Table 1) appear in the main text. In the appendix, "retrofit" denotes the horizon repair of the main text, and "extra@Lk" denotes its late-token perplexity at an LK window (Appendix G); "flatness" refers to low late-token perplexity at long windows, and a "leg" is one trained arm of a comparison. An arm's "transfer tax" is its WT-103 in-window cost, and Δval its change in PG19 validation perplexity. *Pre-registered* (or *registered*) marks a protocol, acceptance band, or decision threshold that was fixed before the corresponding measurement; these criteria are stored with the corresponding result and summary files in the supplementary material.

Table 5: Headline cell and matched controls on the 124M development parent s0. Perplexities are lower-is-better; the needle margin (nats, higher is better) is read in the fixed-grid retrieval battery at its requested depth 0.9, which in these 9.3K-token scenes places the key directly before the question (Appendix G). "Continued" rows share the parent's token stream. GPU-s is measured adaptation wall time on one A100-class device. — marks a cell that was not measured (32K for the learned mixture) or does not apply (GPU-s of the parent).

	In-window PPL		Late-token PPL		Needle	
Arm	PG19 val	WT-103	16K	32K	(nats)	GPU-s
RoPE parent	23.64	327.4	167.6	225.47	−2.73	—
+ matched control (30 updates)	23.64	331.4	168.1	224.19	−4.40	46
+ retrofit, τ=2, own map (30 updates)	23.75	349.2	25.3	26.04	+15.44	74
From scratch, same slope assignment	24.93	604.5	26.7	27.55	+9.49	6212
From scratch, learned mixture	22.84	361.1	24.3	—	+15.64	6750

Table 6: Matched-metric comparison on parent s0. The two cost columns are differences from each arm's baseline: for the 30-update arms (1% dose), ΔWT-103 is taken against the matched 30-update control and ΔPG19 against the untrained parent (the two in-window baselines differ by 0.002 PPL); the 300-update arms (10% dose) use the matched 300-update control for both; training-free and from-scratch arms use the untrained parent. Needle margins (nats) are read in the fixed-grid retrieval battery at its requested depth 0.9, which in these 9.3K-token scenes places the key directly before the question (Appendix G). GPU-s is training wall time; the one-off 48-window horizon measurement is extra for the retrofit arms. — = not measured. †Recommended deployment point. The YaRN protocol is described below the table.

	Cost (ΔPPL)		Late-token PPL		Needle	
Arm	PG19 val	WT-103	16K	32K	(nats)	GPU-s
Retrofit τ=2, own map (1%)	+0.115	+17.81	25.29	26.04	+15.44	74
Retrofit τ=1 (10%)†	+0.005	+0.30	25.01	25.79	+15.80	646
Retrofit τ=2 (10%)	+0.037	+1.27	24.92	25.68	+16.06	623
SelfExtend (w=1024, G=8; 0%)	+0.365	+1.14	27.98	—	+12.81	0
YaRN continuation, YaRN-8 evaluation (1%)	+1.351	+38.25	25.78	52.86	+13.80	53
YaRN checkpoint, plain-head ablation	+1.351	+38.25	173.53	—	−5.54	53
NTK-base rescaling (0%)	+4.888	+51.50	54.19	—	+12.18	0
From scratch, same slope assignment	+1.291	+277.08	26.68	27.55	+9.49	6212

Protocol notes for Table 6. The recommended deployment point (τ=1, 300 updates) has the smallest permanent WT-103 cost of the probed dose ladder (Table 8); the τ=2 300-update row is listed separately because its larger cost (+1.27) distinguishes it from the recommended configuration. The YaRN continuation row reads the continuation checkpoint under the YaRN-8 operator

(YaRN with scale factor 8) on identical windows. Under the pre-registered protocol the YaRN scale factor applies only beyond the 2,048-token training window, so this row's in-window costs coincide with the plain-head reading of the same checkpoint, which is listed separately as an ablation; its needle and 32K cells come from a separate re-probe that applies the YaRN operator consistently. Every arm retains its positional computation at inference.

Table 7: Full public-model comparison on Pythia-1b (native context 2,048; partial rotary). All arms were run in one pass against the same parent evaluation battery; the matched continuation and the horizon repair share token streams. SelfExtend is the two-partition evaluation (w=1024, G=8); "YaRN evaluation" applies YaRN without training, and "YaRN continuation" receives the matched 30 updates inside the YaRN regime. (a) Perplexity, lower is better. (b) Needle cells are content negative log-likelihood (NLL) margins in nats at the fully realized scene length; decode cells are greedy exact-retrieval accuracies over n=64 scenes for five-digit passkeys and word keys. The parent's decode accuracy is at floor (0.00 everywhere), so decode accuracy orders the arms that have signal and supports no claim of absolute capability at this scale.

(a) Perplexity	In-window PPL		Late-token PPL		
Arm	PG19	WT-103	8K	16K	32K
Parent	14.233	12.370	839.3	758.2	777.9
Matched continuation	14.648	12.731	773.0	712.3	725.9
Horizon repair	14.746	13.059	149.0	144.0	127.2
SelfExtend	16.275	13.228	20.7	27.1	125.5
YaRN evaluation	56.198	61.901	181.0	238.2	500.2
YaRN continuation	15.317	14.206	16.1	17.3	93.8

(b) Retrieval	Needle margin (nats)		Decode accuracy (fraction)	
Arm	4K, depth 0.9	16K, depth 0.9	Passkey, 4K	Word key, 16K
Parent	−6.36	−1.36	0.00	0.00
Matched continuation	−6.78	−1.83	0.00	0.00
Horizon repair	+4.31	−1.68	0.00	0.00
SelfExtend	+16.50	+15.75	0.91	0.98
YaRN evaluation	+1.37	+0.53	0.00	0.00
YaRN continuation	+16.53	+14.58	0.77	0.78

B MAP CONTROLS AND DEPLOYMENT REACH

Three controls decompose the recipe's benefit at the main scale. *Coverage*: a procedural ladder holding the map's within-layer structure but leaving the 300–1,200-token band bare (coverage $C = 0.167$ against 0.708–0.77) loses flatness robustly — extra@16k 33.5–46.4 across the three seeds against 25.3–25.4 for the measured maps, monotone at each of the nine measured cross-pairs — while its addressing recovery is seed-dependent: needle margin at depth 0.9 of +15.6–+16.1 on the second and third seeds (within 1.7 nats of the measured-map legs' +17.3–+17.5) yet only +3.9 on the first (against +15.4). *Value multiset*: redrawing each layer's twelve map values i.i.d. from the layer's measured empirical distribution (sampling with replacement, so realized multiplicities are perturbed) pays a sharply larger transfer tax (WT-103 +50.4 against the multiset-preserving shuffle of the same map, whose own tax is only +13.9). The measured *values*, jointly with coverage, therefore keep the transfer tax low, but they are not what carries the deep-depth recovery: identity-free draw-downs of the same band (per-layer redraw, tail-matched, strict-tail, iterated; n=16 paired scenes per probe) preserve it almost completely (+15.8–+16.9 against +15.4 at depth 0.9), while scattering widely at mid-depth (+1.2 to +7.4 against +6.7 at depth 0.5). *Identity*: an aggregate-identical shuffle reproduces flatness *and* addressing; what per-head identity buys is only lower in-window cost (+0.12 vs. +0.21 validation). We report the corresponding negative result: a donor-strength rule over map statistics formalizing "strong maps rescue, weak maps fail" is confounded between its two ingredients at our sample of three distinct parents, both attempted de-confounding routes failed their pre-registered bands, and we report it as an open decomposition rather than a

law. The records of both routes and their pre-registered criteria are included in the supplementary material.

Calibration-free schedules. A sharper substitute question asks whether the horizons must be *measured* at all, or whether a fixed per-head slope schedule at the same dose suffices. Three calibration-free schedules, each adapted for 30 updates at τ=2 on each of the three parents through the identical stream, answer it (Table 3): (i) a procedural geometric ladder from 64 to 16,384 (no diagnostics); (ii) the canonical ALiBi slope set (Press et al., 2021) realized through the same additive operator at matched scaling (m_h=τ/H_h); and (iii) a single-scale uniform schedule (every head H_h=2,048). The benefit of measuring the horizons has the same sign in every cell. In late-token perplexity, the measured map is better in all nine cells, by $+4.9$ to $+44.3$ perplexity at the 16K window (clearing a pre-registered $+5$ margin in eight cells; the ninth, ALiBi on s2, sits at $+4.89$) and by $+5.1$ to $+51.4$ at 32K. For the needle margin, the benefit is likewise positive in all nine cells, $+1.3$ to $+18.5$ nats, but schedule- and seed-dependent in magnitude: the ladder comes within $+1.4$–$+1.7$ nats on s1/s2 yet loses $+11.5$ nats on s0, and the uniform schedule loses everywhere by $+10.4$–$+18.5$. The schedules fail in characteristically different ways: canonical ALiBi changes in-window statistics so strongly that 30 adaptation updates cannot absorb the shock (Δval $+4.5$–$+6.1$ and WT-103 $+370$–$+821$ relative to the measured map, which itself sits $+0.12$ and $+17.8$ above the matched control), the single-scale schedule is in-window benign (Δval ≈ -0.12 relative to the measured map) yet worst at range (extra@16k $+60.1$–$+69.7$), and the procedural ladder splits the difference at a moderate tax. Under the pre-registered scoring rule no cell reaches parity, so the calibration-free substitute is refuted at this dose on every probed schedule: the increment's sign is uniform while its magnitude is schedule- and seed-dependent (the panel's pre-registered thresholds and verdict are stored in its result file in the supplementary material).

Retrieval evidence. The fixed-grid cross-entropy probe improves from -4.40 to $+15.44$ nats, but its scenes contain only 9,264–9,266 tokens and most positive cells place the key near the end. A narrow decoding result recovers 9–11/32 keys with 4,096 filler tokens and formatted decoys, versus 0/32 for the parent; this does not establish full-length retrieval. A dedicated 16,380-token experiment pairs the same 16 contents across distances and checkpoints (Appendix I). After 30 adaptation updates, the retrofit's cross-entropy margin is $+15.36$ with no post-needle filler and $+7.28$ with 1,051 filler tokens, but becomes -1.12, -1.26, and -1.27 at 2,048, 4,096, and 8,192 filler tokens. At 8,192 it still exceeds the matched control by $+6.33$ nats (pointwise paired 95% interval $[4.90, 7.75]$). Relative improvement over a degraded control therefore coexists with a negative contextual margin. A swapped-key decomposition of the same 16 scenes quantifies the remaining advantage: replacing only the needle's digits with a mismatched key, the retrofit's content selectivity (true minus swapped margin) is $+7.76$ $[6.78, 8.75]$ nats at zero filler and $+5.38$ $[4.79, 5.96]$ at 1,051, but collapses to $+0.11$ $[0.02, 0.20]$ at 2,048 and ≈ 0 beyond, while the advantage that survives content erasure still stands at $+6.32$ $[4.90, 7.75]$ nats at 8,192 — the residual relative advantage at range is entirely non-selective (a wrong key in the scene helps the answer score exactly as much as the true key), not recovered addressing. These are teacher-forced answer-NLL measurements on fixed checkpoints. At 300 adaptation updates, the two measured doses likewise have negative margins at 4,096 and 8,192 filler tokens; neither establishes recovered far-range answer scoring. The 30-update profile's shape replicates on two further parent seeds — positive margin through 1,051 filler tokens on all three seeds and negative from 4,096 on, the 2,048 cell itself negative on s0/s1 and point-positive with its paired-t interval crossing zero on s2 ($+0.41$ nats, Appendix I), with the paired advantage over each seed's own matched control still $+6.33$/$+1.89$/$+2.78$ nats at 8,192 (seeds 0/1/2; Appendix I), and the swapped-key decomposition extends to the same panel: selectivity at 8,192 is ≈ 0 on every seed ($+0.00$/$+0.01$/$+0.00$ nats) while the *erasure-proof* (swap-surviving) part carries the full surviving advantage, so the non-selective character of the residual advantage at range replicates on all three seeds under the pre-registered decision rule; the 300-update arms remain single-seed. These comparisons are not generated-key accuracy.

Deployment reach. The recipe's own coverage gate (Section 5) excludes all probed public $\geq$1B checkpoints (OLMo-2-7B, Qwen3-4B/8B, Phi-3-mini, SmolLM2-1.7B; coverage $C \leq 0.087$ in every case, versus $C \approx 0.77$ on our family), as designed. The one public deployment of the full recipe that was attempted was Pythia-410m (C=0.161, partial rotary), and it failed its pre-registered acceptance bands outright. It was benign on validation ($+0.22$) and direction-consistent on late-token

perplexity (compressing the control's extrapolated collapse $15090 \rightarrow 4968$ at the 16K window, and by $3.8\times$ at 32K), but its absolute late-token perplexity sits $\sim 25\times$ above the pre-registered cap of 200, and the needle band failed as well (needle-margin change -3.588 nats in the fixed-grid scenes requested at 16K versus the required $+2.0$). The recipe therefore does not transfer to that checkpoint as-is. The same cells still satisfy the directional rule of Section 2.1 ($r \leq 0.95$ at all three lengths; Table 1), because the rule predicts a relative reduction of late-token perplexity, not these absolute bands. This negative result bounds the demonstrated scope of the full recipe to the measured 124M family plus do-no-harm evidence at 4B (Qwen3-4B decode battery, Appendix F: retrieval and tracking cells unchanged at capped margins, apart from one replicated 32K reasoning-cell harm; benefit unproven there). Sibling transfer of a measured map without new diagnostics is established only *within* the family (two checkpoint pairs).

C Distance-response diagnostic and the dose ladder

Diagnostic. Each head's distance-response profile is the histogram of post-softmax attention mass over query–key distance $d = i - j$, pooled over 48 held-out PG19 windows of 2048 tokens. Query rows are sampled at stride 4 starting at position 32 to skip the beginning-of-sequence (BOS) and attention-sink region; every sampled query contributes its full causal row (row mass one at every query position) with its natural causal truncation, and no per-query range renormalization is applied: left-boundary queries contribute only short-distance mass, so the pooled histogram is the uniform mixture over sampled queries rather than an average of per-query quantiles. Horizons are quantiles of this pooled distribution; we use p90 throughout, read as the nearest-rank upward integer distance at one-token bin resolution — the first bin whose cumulative share reaches 0.90. Median horizons over-localize when used to set the slopes: the median-assignment from-scratch arm reaches validation 26.24 and loses flatness and the low transfer tax in all four headline cells against the p90-map from-scratch sibling. Because the discrete quantile can vanish when $\geq 90\%$ of a head's pooled mass sits at $d < 1$, the pre-registered floor rule (Appendix G) declares zero-horizon heads measurement-undefined, caps their share of any deployed map at 0.10, and imputes such heads to the layer's nonzero median at map-load time, so the deployed slope τ / H_h is never singular. The flagship maps carry 0 zero entries with an observed minimum horizon of 17 tokens (two entries below 20); the public-checkpoint maps are loaded under the same imputation.

Stability across seeds and budgets. Re-measurement across re-trained parents of the same recipe leaves the sorted horizon profile at correlation 0.998 and coverage in $[0.708, 0.771]$; per-head ranks move weakly (Spearman 0.25–0.31) and individual entries by up to an order of magnitude (p90 factor 9), which is why deployment reads aggregates, not identities. Across budgets the same split holds: on the only other-budget parent available (recorded 78.6M training tokens versus the 786M parents, probed under the identical 48-window p90 protocol) the sorted profile re-measures at 0.986–0.989, band coverage is 0.771 (inside the seed range $[0.708, 0.771]$), and per-head ranks stay weak (0.09–0.49) — the aggregate profile survives a ten-fold budget change, while deployment still relies on aggregates. "Across seeds" in this paragraph always means *re-trained* parents of the same recipe, with the diagnostic re-run per checkpoint; it never means window resampling of a single checkpoint.

Stored values and coverage. Two conventions tie every reported horizon and coverage value to a stored result. First, the stored values are the p90 quantile defined above: in the result files, the field that carries them keeps a legacy name (`median_attended_distance`) for compatibility with earlier tooling, but its contents are p90, and every diagnostic file (family and public checkpoints, produced by the same measurement code) records the statistic, the probe context, window count, and data type in its provenance entry. Second, the coverage C of Section 5 is the share of *nonzero-horizon* heads with $H_h \in [300, 1200]$ tokens, computed directly from the stored diagnostics by the script that generates the paper's numbers: the family map gives C=0.77; the five probed $\geq$1B public maps (OLMo-2-7B, Qwen3-4B/8B, Phi-3-mini, SmolLM2-1.7B) all score $\leq$0.087; Pythia-410m scores 0.161. The coverage gate of §5 uses no fitted numeric threshold: it is exactly this separation — the family maps sit far above every probed out-of-family value — and nothing stronger is claimed for it.

Dose ladder. Table 8 and Figure 4 give the τ-dose series at the two operational doses, all cells computed from the corresponding legs. The transient onset at 30 updates is a saturating sigmoid

over its fitted domain (capacity plateau 68.55, midpoint τ_0=2.29 from the five $\tau\leq 8$ points; the τ=12 transient +108.48 overshoots that fitted plateau, so the sigmoid reads inside its domain, not as an all-domain law); the 300-update residue at $\tau\leq 4$ is that onset shrunk by an *approximately constant* relaxation factor (0.072–0.094 across the four probed $\tau\leq 4$ doses; the τ=8 point relaxes at 0.160 on s0 and 0.182 on the independently pretrained s1 parent — a distinct, systematically slower large-τ class on both measured parents that reaches down at least to τ=6 (0.136, s0), and stays on that plateau at τ=12 (0.148), so "invariant" is an approximation here, not an exact law), with no depth plateau at large τ; one independent-parent check (s1, τ=2) falls at +0.22, inside its pre-registered band.

Table 8: τ-dose series on parent s0. WT-103 ΔPPL is the repair's WT-103 perplexity minus that of the matched pure-continuation control with the same number of updates (30 or 300). Late-token PPL is measured at 16K after 300 updates; the needle margin (nats) is read in the fixed-grid retrieval battery at its requested depth 0.5, which in these 9.3K-token scenes realizes depth 0.883, about 1.07K tokens before the question (Appendix G). τ=6 is the s0 boundary point. The last row repeats τ=8 on the independently trained parent s1, against its own matched controls.

	30 updates	300 updates		
τ	WT-103 ΔPPL	WT-103 ΔPPL	Late-token PPL, 16K	Needle (nats)
1	+3.83	+0.30	25.01	+7.91
1.5	+5.45	+0.49	24.90	+8.68
2	+17.81	+1.27	24.92	+7.80
4	+70.87	+6.64	25.07	+7.52
6	+55.27	+7.50	25.22	+7.68
8	+66.24	+10.63	25.35	+5.89
12	+108.48	+16.02	25.61	+4.36
8, parent s1	+73.54	+13.35	25.45	+5.16

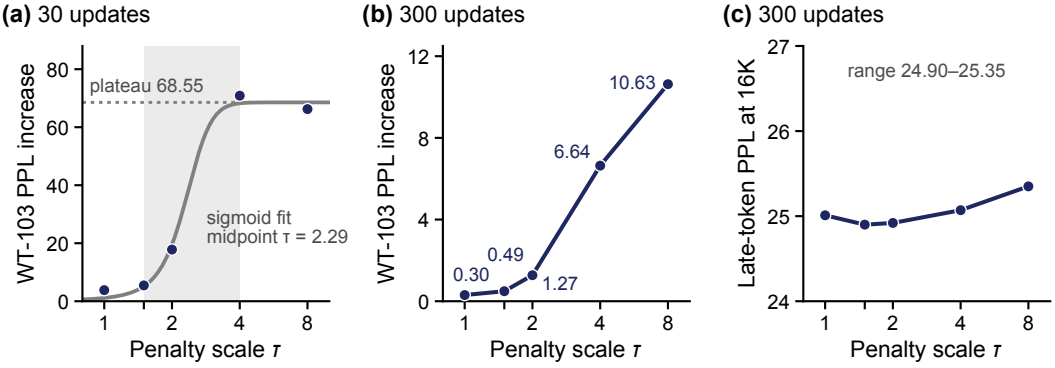


Figure 4: Dose dependence of the horizon repair on parent s0 (log-scaled axis for the penalty scale τ). (a) WT-103 perplexity increase over the matched control after 30 updates: measured points and the sigmoid fit (plateau 68.55, midpoint at τ=2.29). The shaded interval $\tau \in [1.5, 4]$ marks the transition, in which the realized suppression rises gradually and stays below what a sharp (binary) suppression boundary would imply. (b) WT-103 perplexity increase after 300 updates, relative to the matched 300-update continuation; values are printed at each point. (c) Late-token (extrapolated-region) perplexity at 16K after 300 updates; it stays within 24.90–25.35 for every τ (the y-axis does not start at zero). Each point is a single trained run. The five plotted values are $\tau \in \{1, 1.5, 2, 4, 8\}$ (the unlabeled tick is τ=1.5); Table 8 gives their exact values and the additional boundary and independent-parent measurements.

D CROSS-SEED REPLICATION AND MAP TRANSFER

Table 4 in Section 5 lists the cross-seed legs and the two borrowed-map legs. Borrowing the map across sibling checkpoints of the measured family needs no diagnostic run (identity band on every axis, two independent parents); the borrowed legs' residual slack concentrates at mid-depth and is a signal of shared family structure, not identity noise. Cross-*family* transfer was falsified in both directions tested (OLMo-2-7B diagnostic measure: coverage $C = 0.001$, one head of $1{,}024$; Pythia-410m deployment failure, Appendix B).

E SERVING COSTS

Measured batch-1 bf16 inference on the reference PyTorch path (no fused kernels; medians of three repetitions after warmup): steady-state decode tok/s at 4K/16K/32K context is $198.1/194.6/176.8$ (parent), $141.5/137.5/131.0$ (retrofit, exact per-head linear bias), $134.1/138.2/117.9$ (SelfExtend). One-time prefill at 16K/32K: $0.68/2.64$ s (retrofit), $0.79/3.31$ s (SelfExtend), $0.043/0.114$ s (parent). The per-head linear penalty is directly expressible through FlashAttention's native ALiBi-slope interface (per-head slopes; measured fused route at $206.5/200.0/207.5$ tok/s at 4K/16K/32K; layer-level fused-vs-table algebra diff 0.0078; end-to-end fused-vs-reference deviation 0.164, consistent with the measured kernel-family noise floor 0.156 under the pre-registered $1.5\times$ rule (reference-vs-fp64 0.183)); grouped-neighbor attention pays a regrouping pass per forward on this reference path and is likewise available in optimized form in its authors' Triton FlashSelfExtend; the comparison here therefore reports measured reference-path costs, not an availability claim.

F QWEN3-4B DO-NO-HARM BATTERY

On Qwen3-4B-Base (a checkpoint the coverage gate excludes from benefit claims), a 30-update τ=2 capped retrofit (the penalty stays linear near the query but its magnitude is capped at 0.25, $b_h(i,j) = -\min\{\tau(i-j)/H_h,\, 0.25\}$) was run through a held-out decode battery (16 items per cell; retrieval and reasoning items at 4K–32K, exact-match and multiple-choice scoring), and the 32K reasoning cell that decides the verdict was then re-measured on three fresh item draws under the pre-registered replication protocol. Relative to the raw parent, all nine retrieval/tracking cells hold exactly 0.00 in the first draw and in all three fresh draws; the 4K reasoning cell improves consistently ($+43.75$ in the first draw, $+18.75$ to $+37.50$ fresh); the 16K cell reads as draw noise around parity ($+6.25$ first, -25.00 to -12.50 on the fresh draws); and the 32K reasoning cell carries a *stable* harm (-37.50 first, -37.50 to -18.75 on the fresh draws, mean -27.08 against the ± 12.5-point item-draw noise band). The pre-registered replication test therefore rejects the item-draw noise explanation: the residual is a candidate property of the repair, not variation in the sampled items. Cross-entropy guards stay inside band (Δval $+0.004$, ΔWT-103 $+0.012$, extra@16k 10.821 against control 10.814). Whether this harm can be removed remains open. A first amplitude test, a halved-cap variant (0.125), kept all nine retrieval/tracking cells at exactly 0.00 and the cross-entropy guards inside band, yet relaxed the 32K cell only to -25.00, failing its pre-registered band; the residual therefore does not scale with the amplitude cap. It does not track the slope either: at τ=0.75 and τ=0.375 the same 32K cell reads $-56.25/-56.25$ on the pre-registered τ-reduction ladder (mixed verdict). We therefore attribute the far reasoning residual at 4B to the recipe rather than to the dose. Benefit at 4B is explicitly not claimed; the battery is do-no-harm evidence on the retrieval cells plus one named, replicated reasoning cost at 32K.

G CONVENTIONS, CONTROLS, AND EVALUATION PROTOCOL

Units and axes. Perplexity deltas are ppl *points* on the stated evaluation set (validation slice; WT-103; extra@16k/32k). Needle margins are cross-entropy *nats*, never ppl. "extra@Lk" is mean perplexity over positions beyond the 2048 training context in evaluation windows of $L\times 1024$ tokens under the pre-registered protocol; every perplexity in this paper is $\exp$(mean token-level cross-entropy over the stated positions) — never an average of per-window perplexities (equal-length windows, so the window mean coincides with token pooling). Decode battery cells are accuracy deltas in percentage points against the raw parent, n=16 items per cell. Per-experiment window counts (never pooled): distance-response diagnostics 48 held-out PG19

windows at 2048; public-checkpoint map probes 24; validation perplexity 96 PG19 holdout windows at 2048; long-document perplexity up to 24 PG19 test books in the custom family and up to eight sufficiently long books in each public-checkpoint panel; the forward logit-displacement probe (mean squared logit change against the parent; its code is in the supplementary material, and no reported number depends on it) 16 windows$\times$2048 per domain. Intervals marked [95%], "t_{95}", or "t_{15}" are the same pointwise paired 95% t interval over the n=16 paired contents: $\mathrm{mean} \pm t_{0.975}(15)\, s/\sqrt{16}$, s the sample (unbiased) standard deviation, $t_{0.975}(15)$≈2.131. Deltas use unrounded cells and need not match displayed differences.

Retrieval grids. Needle and passkey scenes are token-budget-constructed: the scene budget includes the prompt template and the filler block truncates at its realized end, so the grid point requested at 16K realizes exactly 9,264 tokens under the reference token-count class (filler 9,217 + prefix 17 + needle 15 + question 13 + answer 2; 100 of the 144 sampled five-digit strings across the three measured seeds), and 9,266 under the residual class (needle 16 + answer 3; the other 44 strings); every requested depth $\geq$0.5643 collapses to key-start token 9,234 in both classes (realized depth 0.9968; 0.9965 in the longer class), with the key 30 (32) tokens from the scene end and zero filler between key and query, and the requested 0.5 cells realize 0.883 with a 1,066–1,068-token margin. All labels from this fixed-grid battery inherit this realization; the dedicated distance experiment in Appendix I instead uses complete 16,380-token scenes. The fixed token-position lists of every retrieval scene in this paper — the 288 fixed-grid needle scenes, the 180 passkey decode scenes, and the 80 shared paired-far layouts — are included in the supplementary material and were checked sample by sample against the stored offset records of all eighteen paired-far result files. Because every leg of a given seed shares the identical digit stream, this class mixture is itself identical across arms and cancels in any cross-arm difference; a post-hoc split of every per-sample margin cell by digit class (script and output in the supplementary material) confirms no class carries the recovery: every registered retrofit leg stays positive within each class separately at the requested 0.5 depth, the small within-arm class gap flips sign across depths, and the registered controls sit at floor in both classes. Probe provenance: at every (arm, depth), the seven arms of an earlier probe version re-evaluate within 0.11–1.05 nats of their recorded values under the current construction, with the drift's sign covarying with the deployment-map path; a pre-registered two-stage, eight-variant bisection matched none of the earlier constructions, so the drift reflects an unidentified model-side revision. Reported values are the recorded ones and are never blended with re-evaluations. Labeling cells by their realized length changes no reported number. The swapped-needle control permutes the fact-key content while preserving scene structure; it quantifies the selective (≈+9 nats) versus non-selective residue of the margin recovery. Passkey exact-retrieval accuracy is at floor across the battery: every recorded passkey cell is exactly zero except two cells of the scenes requested at 16K (depths 0.7 and 0.9) of one ablation arm (the ALiBi mapping on the s2 parent), each carrying a single correct sample out of three; it is reported as such: the raw decode signal is capability-limited and carries no positional information here (checked by an audit script included in the supplementary material).

Comparator settings. SelfExtend: window w=1024, group G=8, 12 layers patched, our implementation checked algebraically against an fp64 per-pair reference inside the real module. YaRN continuation: train-time scale-8 for the matched 30 updates and evaluated under the matching YaRN-8 regime (25.78) and under the plain head (173.53); NTK-base rescaling is evaluated at zero compute ($\theta' = 10^4 \cdot 8^{64/62}$). Matched continuation controls share token streams, seeds, and remaining schedule with their retrofit legs. All from-scratch arms train 786M tokens at context 2048.

Horizon maps. Deployed slope maps use the p90 distance-response statistic of Appendix C. A head whose discrete p90 vanishes is measurement-undefined: such heads are capped at 0.10 of any deployed map, excluded from the map's descriptive statistics, and imputed to the layer's nonzero median at load time, so the deployed slope τ/H_h is never singular.

H SCORE-LEVEL MECHANISM STATEMENTS

This appendix states, at score level and under idealized content, what connects the two channels of equation 2; it makes no task-level or optimization claim. Write a head's pre-softmax score at lag

δ as $g_{q,k}(\delta) - \frac{\tau}{h}\,\delta$, where $h = H_h$ is the head's horizon, $g_{q,k}(\delta) = d_h^{-1/2} \sum_{j=0}^{D-1} \big[a_j \cos(\omega_j \delta) + b_j \sin(\omega_j \delta)\big]$, $\omega_j = \theta^{-2j/d_h}$, $D = d_h/2$, and $a_j = q_{2j}k_{2j} + q_{2j+1}k_{2j+1}$, $b_j = q_{2j}k_{2j+1} - q_{2j+1}k_{2j}$ collect the per-pair content products.

Exact statements (independent content). For q, k with i.i.d. $\mathcal{N}(0, \varsigma^2)$ coordinates and distinct keys independent, with $\gamma(\Delta) := D^{-1} \sum_j \cos(\omega_j \Delta)$: (i) $\mathbb{E}\, g(\delta) = 0$, $\mathrm{Var}\, g(\delta) = \varsigma^4$, and $\mathbb{E}[g(\delta)g(\delta')] = \varsigma^4 \gamma(\delta - \delta')$: the phase channel is a stationary cosine-sum kernel. (ii) The gap contrast $\sigma^2_{\mathrm{ph}}(\Delta) := \mathrm{Var}\big(g(\delta{+}\Delta) - g(\delta)\big) = 2\varsigma^4\big(1 - \gamma(\Delta)\big)$ is strictly positive on $0 < \Delta < \Lambda$, since $\gamma(\Delta) = 1$ requires every cosine to be 1, in particular $\cos(\omega_{\min}\Delta) = 1$; at our $d_h{=}64$, $\theta{=}10^4$ the first wrap sits at $\Lambda = 2\pi\theta^{(d_h-2)/d_h} = 47{,}117$ tokens — beyond the longest realized scene in this paper (16,380) — and the unit-gap contrast evaluates to $\sigma_{\mathrm{ph}}(1) = 0.260\,\varsigma^2$ (exact sum at the stated base). (iii) The margin term contributes the deterministic, content-free difference $M = \tau\Delta/h$ at every gap: a recency prior of strength independent of content. The channels are complementary by construction — one is content-dependent with oscillatory stationary covariance (positivity of (ii)'s contrast alone does not imply that this channel keeps one direction all the way to the wrap; γ itself oscillates); the other is content-free and uniform over distance.

Aligned content: non-stationary contrast. For a retrieval-aligned key $k^\star = c\,q + \varsigma\sqrt{1-c^2}\,z$ ($c \in [0,1]$, an effective statistic produced by training, here assumed), the per-pair moments are $\mathbb{E}\, a_j = 2c\varsigma^2$, $\mathbb{E}\, b_j = 0$, $\mathrm{Var}(a_j) = 2\varsigma^4(1+c^2)$, $\mathrm{Var}(b_j) = 2\varsigma^4(1-c^2)$ — the edge case $c{=}1$ forces $b \equiv 0$ deterministically — and

$$\begin{aligned}\mathrm{Var}\big(g(\delta{+}\Delta) - g(\delta)\big) &= \frac{4\varsigma^4}{d_h} \sum_j (1 - \cos\omega_j\Delta)\,\big(1 - c^2 \cos\big(\omega_j(2\delta{+}\Delta)\big)\big) \\ &\in \big[(1-c^2),\,(1+c^2)\big]\,\sigma^2_{\mathrm{ph}}(\Delta). \end{aligned} \tag{3}$$

The upper end of this bracket is necessary: at $c{=}1$, $\Delta{=}\pi$ the left side evaluates to 0.576 against the factor-free value $\sigma^2_{\mathrm{ph}} = 0.420$ (computed in closed form and verified numerically against sampling). Let $X = g(\delta_1) - g(\delta_2) + M$ have mean $\mu = S_{\mathrm{ph}} + M$ and variance V, where $S_{\mathrm{ph}} = c\varsigma^2\sqrt{d_h}[\gamma(\delta_1) - \gamma(\delta_2)]$. For $\mu > 0$, Cantelli's inequality gives the distribution-free moment bound

$$\Pr(X \le 0) \le \frac{V}{V + \mu^2}.$$

Under a Gaussian approximation one instead obtains $\Pr(X \le 0) \approx \Phi(-\mu/\sqrt{V})$. If $S_{\mathrm{ph}} \ge 0$ and $M \ge 0$, the variance bracket implies that this Gaussian approximation is at most

$$\Phi\left(-\frac{\max(S_{\mathrm{ph}}, M)}{\sqrt{1+c^2}\,\sigma_{\mathrm{ph}}(\Delta)}\right).$$

This last expression is an upper bound on the Gaussian approximation, not an exact reversal probability for Gaussian-product scores. The covariance $\varsigma^4\gamma(\delta - \delta')$ entering this bound is the covariance of *one* key's score at two counterfactual placements; between two distinct independent keys sharing the same query the cross-covariance vanishes, so a two-key ranking variance would be $2\varsigma^4$ per unit gap rather than $2\varsigma^4\big(1 - \gamma(1)\big)$, and nothing in this paper scores that two-key object.

Scope of the score-level statements. The phase channel supplies a content-dependent tie-break of typical size $\varsigma^2\sigma_{\mathrm{ph}}(1)$ per unit gap; the margin supplies a deterministic recency prior that also *opposes* a matched key when a nearer competitor exists — the operator is a prior, not a retrieval guarantee, which is consistent with the far-distance decay measured in Appendix I. The crossover distance at which the phase channel's per-token residual rate falls below τ/h can be written in closed form only through an envelope heuristic: under a retained base (the deployed arm, all heads at $\theta{=}10^4$), $\delta^* = h\,\kappa\,W_0(\rho/\kappa)$ with $\kappa = \mu_0/(\tau \ln\theta)$ and $\rho = 2\pi\theta/h$, giving median $\delta^*/h = 1.83$ on the measured 144-head map with per-head alignment fits μ_0 (median $R^2 = 0.93$) — materially different from the 0.548 of the re-based convention, and in any case not a quantity that a costless threshold mask can identify (a null at one threshold is consistent with every smaller one). We report the crossover as a heuristic scale only; the complementarity claim above rests on the exact constituents, and no statement here predicts benchmark behavior. The numerical examples use the exact finite frequency sums at the stated base; they describe an idealized score distribution, not trained-task accuracy.

I CONTENT-PAIRED FULL-LENGTH DISTANCE MEASUREMENT

The scene builder for this experiment supplies enough filler to realize exactly 16,380 tokens at every requested distance. It varies the filler after the needle over $0, 1051, 2048, 4096, 8192$ tokens, with the fixed query and answer following it. Thus the post value measures filler length, not the entire key-to-answer token separation. Sixteen independently generated five-digit keys and filler offsets are reused across all five distances and all nine checkpoints by resetting the same RNG at each distance. The first pair uses the 30-update retrofit and its matched continuation control. The extension adds 300-update retrofits at two doses and their matched continuation control: all five of these checkpoint evaluations are from the same parent training seed, and the 30-update pair's shape is then replicated on two further parent seeds below. The margin is the summed answer NLL under the question-only input minus that under the full scene. Positive margin means the scene improves this answer score; it does not directly measure decoding success.

Post filler	Retrofit margin	Control margin	Paired difference [95%]
0	15.356	−4.264	19.620 [18.074, 21.166]
1051	7.281	−7.612	14.893 [12.789, 16.996]
2048	−1.122	−7.642	6.519 [5.034, 8.005]
4096	−1.265	−7.656	6.391 [4.998, 7.785]
8192	−1.265	−7.590	6.325 [4.902, 7.748]

At 8,192 filler tokens the method advantage drops by 13.294 nats from post 0 (paired difference-in-differences interval $[-14.208, -12.381]$). Intervals use a pointwise t_{15} reference for paired scene differences; they are not simultaneous bands or uncertainty across training seeds. The fixed-grid short-scene battery uses a different content-sampling definition and is reported separately.

300-update extension. With 8,192 filler tokens, the 300-update $\tau = 2$ and $\tau = 1$ margins are -0.796 and -0.843 nats versus the matched control's -6.988. Their paired advantages are 6.192 $[4.779, 7.605]$ and 6.146 $[4.771, 7.520]$. Thus additional adaptation retains a relative advantage while leaving the absolute contextual margin negative. Across post filler lengths $0/1051/2048/4096/8192$, $\tau = 2$ margins are $16.190/8.692/-0.272/-0.801/-0.796$ and $\tau = 1$ margins are $15.937/8.754/0.287/-0.831/-0.843$. The small positive $\tau = 1$ mean at 2,048 does not by itself establish a reliably positive margin; its pointwise interval includes zero. All intervals use the same 16 paired contents. The five checkpoint evaluations support distance sensitivity of the measured contextual advantage, without expanding the claim to successful far-range decoding.

Cross-seed shape replication. The 30-update pair's profile rests on one training seed above. Repeating the identical instrument — same 16 contents, grid, metric, and protocol — on the two independent parent seeds' stored 30-update checkpoints (own-map τ=2 retrofit and matched continuation control each) gives a three-seed, sixteen-scene paired panel:

	Own margin			Paired own−control [95%]		
Post filler	s0	s1	s2	s0	s1	s2
0	15.356	15.258	16.784	19.62 [18.07, 21.17]	16.20 [14.94, 17.47]	20.21 [18.63, 21.80]
1051	7.281	9.414	8.657	14.89 [12.79, 17.00]	11.36 [10.01, 12.71]	12.52 [10.72, 14.31]
2048	−1.122	−0.202	0.409	6.52 [5.03, 8.00]	1.76 [0.75, 2.77]	4.28 [2.91, 5.64]
4096	−1.265	−0.271	−1.124	6.39 [5.00, 7.78]	1.75 [0.77, 2.73]	2.75 [1.66, 3.85]
8192	−1.265	−0.263	−1.122	6.33 [4.90, 7.75]	1.89 [0.91, 2.87]	2.78 [1.69, 3.87]

On all three seeds the same qualitative pattern holds: positive own margins at posts 0/1,051, negative point estimates from 4,096, and a paired advantage over each seed's matched control staying positive at 8,192 (per scene: positive in 46 of the 48 scene×seed pairs from post 2,048 on, 48 of 48 at posts 0/1,051) — the shape replicates in sign on three training seeds, with seed-level *magnitudes* differing descriptively (the s0 control degrades most at range, -7.59 against -2.15/-3.90 nats at 8,192 on s1/s2; on s2 the 2,048 own cell is point-positive at $+0.41$, its interval crossing zero, like the single-seed τ=1 cell above). The 300-update arms of the previous paragraph were measured on s0 only. Pooling the 48 own-arm margins per far post, a scene×seed decomposition attributes 34–42% of the

margin variance to scenes, 11–17% to the training seed, and 47–49% to interaction, and no single scene carries more than 0.37 of any far-post cell's dispersion; in the swap battery all sixteen scenes have strictly positive selectivity at posts 0/1,051 on all three seeds.

Swapped-key decomposition of the surviving advantage. The distance profile above leaves two readings of the surviving relative advantage open: content-selective recovery of the true key at range, or non-selective suppression that would help the answer score even with the key's content erased. A second measurement on the same 16 scenes decides between them. An independent stream replaces only the needle's five digits with a mismatched key (question and answer stay the true key); scenes, offsets, and refusals are paired across distances and checkpoints, with zero refusals realized. Per scene, the retrofit's content selectivity is $\mathrm{SEL} = m_{\mathrm{true}} - m_{\mathrm{swap}}$ and the advantage that survives content erasure is $\mathrm{NONSEL} = m_{\mathrm{own,swap}} - m_{\mathrm{ctl,true}}$; the exact identity $\mathrm{own}_{\mathrm{true}} - \mathrm{ctl}_{\mathrm{true}} = \mathrm{SEL}+\mathrm{NONSEL}$ holds to 10^{-6} per cell, the true-key NLLs recomputed in the swap probe reproduce the recorded margins (max cell-mean deviation < 0.02), and the pre-registered selectivity threshold is $T_0 = +2.0$ nats with pointwise paired t_{15} intervals.

SEL (selective part)			
Post filler	s0	s1	s2
0	7.76 [6.78, 8.75]	7.29 [6.30, 8.29]	8.18 [7.22, 9.13]
1051	5.38 [4.79, 5.96]	6.89 [6.11, 7.66]	5.90 [5.25, 6.55]
2048	0.11 [0.02, 0.20]	0.04 [0.01, 0.06]	1.11 [0.29, 1.92]
4096	0.00 [−0.01, 0.01]	−0.00 [−0.01, 0.01]	−0.00 [−0.01, 0.01]
8192	0.00 [−0.01, 0.01]	0.01 [−0.01, 0.02]	0.00 [−0.01, 0.01]

NONSEL (erasure-proof part)			
Post filler	s0	s1	s2
0	11.86 [10.01, 13.71]	8.91 [7.30, 10.52]	12.04 [10.25, 13.82]
1051	9.52 [7.46, 11.57]	4.47 [3.01, 5.94]	6.62 [4.90, 8.34]
2048	6.41 [5.00, 7.82]	1.72 [0.72, 2.72]	3.17 [1.87, 4.47]
4096	6.39 [5.00, 7.78]	1.75 [0.77, 2.74]	2.75 [1.66, 3.85]
8192	6.32 [4.90, 7.75]	1.88 [0.90, 2.86]	2.78 [1.70, 3.87]

For the 30-update pair, selectivity clears the pre-registered threshold at posts 0 and 1,051 and falls to $\leq$ 0.12 nats from 2,048 onward, where the entire residual advantage is erasure-proof — and this verdict replicates on both independent parent seeds (the pre-registered cross-seed verdict: each seed's selectivity is classified SELECTIVE at posts 0/1,051 and NONSELECTIVE at 8,192 under the pre-registered threshold $T_0 = +2.0$). On s2 the 2,048 selectivity cell is point-positive at 1.109 $[0.294, 1.923]$ — its interval excludes zero yet sits below the pre-registered threshold, so the rule classifies it NONSELECTIVE; from 4,096 onward selectivity is $\approx$0 with intervals containing zero on both additional seeds, while the erasure-proof part carries the full surviving advantage ($+6.32$/$+1.88$/$+2.78$ nats at 8,192 on s0/s1/s2). Both 300-update pairs (s0 only) show the same verdict: τ=2 selectivity 8.182 $[7.115, 9.249]$ at post 0 decays to 0.437 $[0.153, 0.721]$ at 2,048 and $\approx$0 beyond, with 6.189 $[4.775, 7.603]$ erasure-proof at 8,192; τ=1 selectivity 8.154 $[7.110, 9.198]$ decays to 0.928 $[0.313, 1.544]$ at 2,048 and $\approx$0 beyond, with 6.146 $[4.776, 7.514]$ erasure-proof at 8,192. The matched controls' own selectivity is negligible everywhere (≤ 0.46 nats at every distance on s0; ≤ 0.16 and ≤ 0.03 on s1/s2), so the near-end selective signal is an operator-induced addressing effect, and its disappearance at long distances is shared. The first row also decomposes the near end: of the 19.62-nat s0 advantage at zero filler, 11.86 is already content-blind. The mechanism that survives to 8,192 filler tokens is therefore, on all three seeds, a symmetric scene-level suppression, not recovered content-addressing at range.

Fixed-grid scene battery. Verbatim far-key (passkey) decoding is at floor for every arm at this scale, including the parent (two isolated single-sample decodes in one s2 ALiBi cell pair aside, Appendix G), so the raw decode signal carries no positional information here. The addressing benefit is therefore reported on the cross-entropy needle probe: margin $-4.40 \rightarrow +15.44$ nats with a disclosed non-selective residue (selective component $\approx +9$ nats; span-selectivity share 0.130, i.e., the fraction of the repair's added attention mass on the key span, relative to the control, that is lost when

the key is swapped; below the 0.40 pre-registered bound, i.e. *refuted* as span-selective addressing at the deployed dose). A narrow decode-level positive exists: with 4096 tokens of filler plus formatted decoys, the parent emits no digits ($0/32$) while the retrofit resolves the true key 9–$11/32$; far-range decoding and 2-digit readout nonetheless remain unsuccessful, and the scenes requested at 16K realize $9{,}264$ tokens ($9{,}266$ for the 44 longer-digit scenes, Appendix G) with the key standing 30 (32) tokens from the scene end for every requested depth ≥ 0.5643 and a $1{,}066$–$1{,}068$-token margin at the requested 0.5 depth (all retrieval labels carry this realization; exact offsets from the scene builder), so every positive needle cell in that fixed-grid battery is near-end cross-entropy recovery under long distractor context, not far-distance key–query separation. Depth-resolved within the same fixed battery, the requested-0.5 cells sit at $\approx 1.07 \times 10^3$-token key–query separation and are positive on every retrofit dose leg — each leg's 16-sample t_{95} interval sits strictly above zero, checked per leg by an audit script — ($+6.71$ nats at the headline dose; the 10%-dose family reads $+7.80$–$+10.25$ across the three parents at τ=2 and $+4.36$–$+8.68$ across its full τ ladder, against registered matched controls of -6.36 (s0) and -1.72 (s1) — the s2 300-update control's needle cells were never measured, and values from the earlier protocol are not quoted), while the requested-0.1 cells ($\approx 7.6 \times 10^3$-token separation) read -2.03–$+0.27$ across the measured dose legs and parents (-7.23/-2.33 for the same two registered controls) — the positive addressing evidence reaches about a ninth of the realized scene and decays to annihilated-but-not-recovered at its far end; the fixed-grid battery contained no full-length far-distance cell. Nothing in this paper claims demonstrated long-context retrieval; the claim is cross-entropy-level recovery with a quantified proxy gap (Liu et al., 2023; Hsieh et al., 2024; Akbar, 2026).

J Development and prospective evidence

The public diagnostic panel (Table 1, Figure 3) distinguishes development observations from predictions fixed before their outcomes. OLMo-2-7B-Instruct, Pythia-1b, and Pythia-2.8b are the three prospective checkpoints on which the rule fires; Qwen3-0.6B and Qwen3-1.7B were also selected prospectively, and the rule is silent on them. OLMo-2-7B, TinyLlama-1.1B, and Phi-3-mini-4k were measured under an earlier scoring rule, which used the same trigger ($B \geq 2$) but predicted a larger reduction ($r \leq 0.70$); their cells are listed under the current rule as earlier evidence, not as fresh prospective tests; Pythia-410m and SmolLM2-1.7B were used for development.

K Extended discussion of limitations

This section gives the detail behind the Limitations and future work paragraph of Section 7.

Scope of the public evidence. The separation is not a population-level guarantee or a calibrated probability for a new model. Neither checkpoint selection nor repeated lengths supplies random sampling from a model population: lengths within a checkpoint are correlated, only five checkpoints (three fired, two silent) were selected prospectively, and the earlier and development checkpoints are not fresh validation. The panel does not locate the threshold precisely; the rule mostly separates clearly collapsing checkpoints from nearly flat native-long-context ones. With the split fixed at 2,048 tokens, "late" is not always beyond a model's pretraining window, and B need not be invariant across corpora; B requires training the matched continuation, so it is not a zero-training property of the weights. A silent rule makes no prediction: it neither selects the best repair nor implies that a silent model would fail to benefit from every intervention. Two stronger hypotheses, a two-axis dose-geometry rule and a magnitude prediction of $r \leq 0.70$, failed and are not supported by the present results. Perplexity cells carry no uncertainty intervals (three training seeds; single-run public cells), and retrieval intervals cover 16 contents at fixed weights, not training seeds.

Method choice. A ratio $B \geq 2$ mainly signals deterioration after 2,048 tokens, so the rule detects need rather than anything repair-specific. The repair is not the best method at range: on Pythia-1b, SelfExtend (27.1 at 16K, no adaptation) and matched YaRN (17.3) reach far lower late-token perplexity than the repair (144.0); on the development model, SelfExtend avoids adaptation and the learned-mixture baseline (trained from scratch on the full 786M tokens; Table 5) remains the better in-window model. The alternatives were run only on this one public checkpoint and the development family, so we cannot test whether the rule discriminates among methods or whether they would also

fail to help on silent cells. The rule does not show that this repair is the right response to the need, and we make no superiority claim over neighboring methods without matched implementations.

Costs and reach. The cross-domain cost is much less stable than the in-domain cost ($+17.8/+1.4/+5.7$ WT-103 points). The 300-update $\tau = 1$ result ($+0.30$) is one seed-0 dose selected from its ladder, not a prospective test of the 30-update rule or evidence that transfer costs always disappear, and larger τ values relax more slowly (Appendix C). GPU-seconds exclude pre-training and the measurement sweep, and decode speeds are reference-path measurements (Self-Extend also has an optimized kernel). The headline repair is seed-level matched continuation on one 124M family at one operational point, and sibling-map transfer is established only within that family. Low-coverage checkpoints lie outside the full-outcome claim even where the rule holds, and the coverage gate has no fitted numeric threshold. On Pythia-410m, the one public checkpoint where the full recipe was attempted as a deployment, the cells satisfy the rule but the recipe failed its pre-registered absolute bands (16K late-token perplexity $15090 \rightarrow 4968$ against a cap of 200, about $25\times$ the cap; needle-margin change -3.588 nats against a required $+2.0$; Appendix B): the rule predicts a relative reduction, not a usable long-context model. On Qwen3-4B-Base the repair repeatedly harmed the 32K reasoning cell (-37.50 points in the first draw, -37.50 to -18.75 on three fresh draws of 16 items), and neither halving the cap nor lowering τ removed that harm, so at 4B we report do-no-harm evidence on retrieval only, not benefit (Appendix F). A *donor-quality predictor*, which would predict from a donor map's statistics whether the map rescues a recipient, remains confounded between value distribution and coverage, and the calibration-free comparison covers three schedules, not an exhaustive search.

Retrieval and mechanism. A distance penalty is not a retrieval mechanism: as a recency prior it may improve aggregate prediction while suppressing a useful distant key. In the 16,380-token experiment, the seed-0 repair's content selectivity falls to $+0.11$ nats at 2,048 filler tokens and about zero at 8,192, where a wrong key confers nearly all of its $+6.33$-nat advantage over the degraded control; the pattern repeats across parents, so lower loss does not restore distant-key retrieval (Appendix I). These are teacher-forced loss contrasts, not generated-key accuracy: verbatim far-key decoding is at floor across the fixed-grid battery (Appendix G), public-model decoding gives the stronger far-content signal to SelfExtend and YaRN (Table 7), and the fixed-grid "16K" scenes held only about 9.3K tokens, with many keys near the end. A loss reduction that survives a key swap is not evidence of content retrieval, and neither the score-level algebra, a lower loss, nor passing probes establishes a generally safe long-context repair. The rotary-base observation neither identifies a unique site of positional memory nor invalidates well-adapted phase-scaling methods, and our controls give no theory of how head horizons emerge in training. Broader deployment would require additional corpora, independent model families, and task outcomes.