Mechanics of Long-Context Hybrid Models

Part 1.1: From Hybrid Attention to Hybrid Position
Research group reading: mechanisms, interventions, and evidence

Paper authors: Xiaoran Liu, Ziwei He, Xipeng Qiu

arXiv:2610.10114v1 | October 10, 2026

Local positional control and global retrieval need different recipes
for inference extrapolation and continued training.

Two ways to obtain a longer context

Length extrapolation

Train at 4K, then evaluate at 8K–64K.

No weight updates.

Question: can the learned computation survive a longer input?

Context extension

Continue training at 32K for 5B tokens.

Weights are updated.

Question: does the model learn to use additional distant context?

64K remains beyond the training length even after 32K extension.

Paper, Secs. 1 and 3.1, PDF pp. 2–4.

A controlled comparison, with two training stages

	Initial pretraining	Continual pretraining
Context length	4K	32K
Training tokens	50B	5B
Corpus	DCLM-Baseline-1.0	LongWanjuan
Batch tokens	1M	0.5M
Peak learning rate	10^{-3}	10^{-3}

- Model sizes: 376M, 776M, 1B; additional 3B verification.
- MHA; Llama 3 tokenizer; default efficient:NoPE ratio = 3:1.
- Retrieval: RULER. State tracking: BABILong. Modeling: PG19.

Paper, Sec. 3.1, p. 4; Appendix A and Table 2, p. 32. Length and corpus both change across stages.

The global branch is paired with positional bias

Branch	Direct access	Positional structure
NoPE full attention	All past tokens	No explicit RoPE
RoPE full attention	All past tokens	Rotary positions
SWA	Recent window	RoPE + local mask
GLA / GDN	Recurrent state	Data-dependent gates

Layer-wise hybrid (LH)

Three efficient layers, then one full NoPE layer, repeated.

Head-wise hybrid (HH)

Efficient and full NoPE heads collaborate within each layer.

Paper, Sec. 3.1, p. 4. Default SWA window: 128. Full NoPE still stores a length-dependent KV cache.

A recurrent gate is an implicit position mechanism

For GLA, an interaction between positions s and t contains

$$q_t \left(\prod_{i=s+1}^{t} \mathrm{Diag}(\boldsymbol{\alpha}_i) \right) k_s^\top .$$

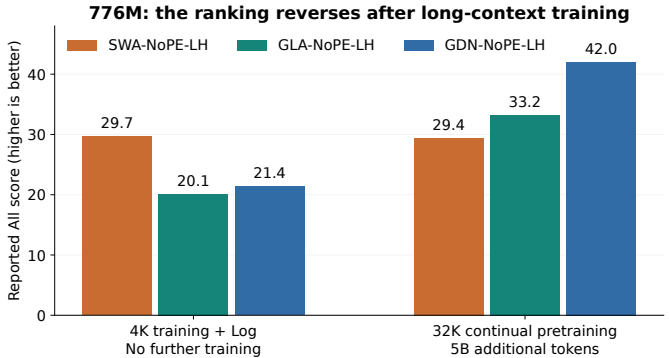
- q_t, k_s: query and key; $\boldsymbol{\alpha}_i$: data-dependent gate.
- Distance changes how many gate factors are accumulated.
- No explicit position embedding does not imply no position bias.

Teaching example: a constant scalar gate

$0.99^{100} \approx 0.366$, while $0.99^{1000} \approx 4.3 \times 10^{-5}$.

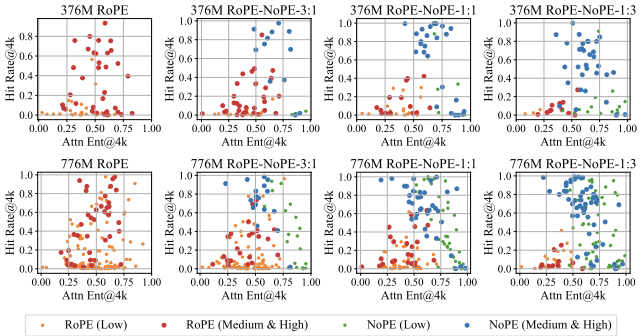
Paper, Sec. 6.2, p. 21. Scalar numerical example is an explanatory simplification.

The SWA advantage reverses after continued training



Redrawn from Tables 7 and 11, pp. 43 and 47. 776M, 3:1 LH. All averages 12 benchmark–length cells.

NoPE heads locate globally but often aggregate broadly



Each point is a head: entropy on the horizontal axis, top-64 hit rate on the vertical axis. Rows: 376M / 776M.

Original Fig. 12(b), p. 13. Columns: pure RoPE, then RoPE:NoPE = 3:1, 1:1, 1:3. Colors also encode retrieval-score bins.

The best mixture depends on the attention family

776M HH; reported All score	3:1	1:1	1:3
RoPE-NoPE + EME	**30.4**	25.7	22.3
SWA-NoPE + Log	31.6	**35.2**	31.9

- More global heads do not give monotonically better results.
- In 1:1 HH at 4K, perturbing positional heads hurts more than perturbing NoPE heads.
- Authors' interpretation: global coarse localization needs concentrated processing by other heads.

Paper, Sec. 4, Figs. 10–13, pp. 11–13; Table 17, p. 52. Ratios are positional:NoPE.

A Jacobian makes attention concentration comparable

For softmax attention, holding queries and keys fixed:

$$o_t = \sum_s \alpha_{t,s} v_s, \qquad J_{t,s} = \frac{\partial o_t}{\partial v_s} = \alpha_{t,s} I, \qquad \|J_{t,s}\|_2 = \alpha_{t,s}.$$

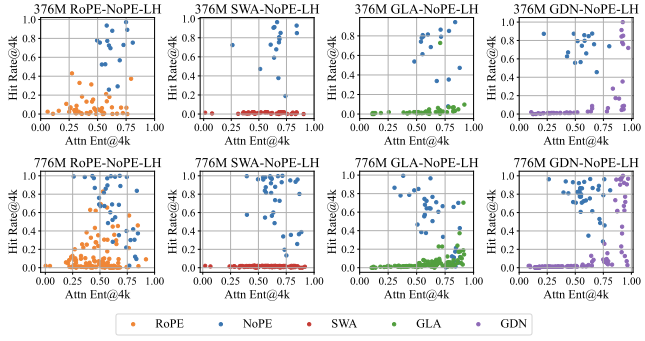
Normalize the same nonnegative quantity for other attention operators:

$$\widetilde{\alpha}_{t,s} = \frac{\|J_{t,s}\|_2}{\sum_u \|J_{t,u}\|_2}, \qquad \widetilde{H}_t = -\frac{\sum_s \widetilde{\alpha}_{t,s} \log \widetilde{\alpha}_{t,s}}{\log L}.$$

- $\|\cdot\|_2$: spectral norm; L: context length; nonzero denominator.
- Recovers softmax attention weights without changing LA computation.

Paper, Eq. (1), Sec. 6.1, p. 18. Normalization is written out from the accompanying description.

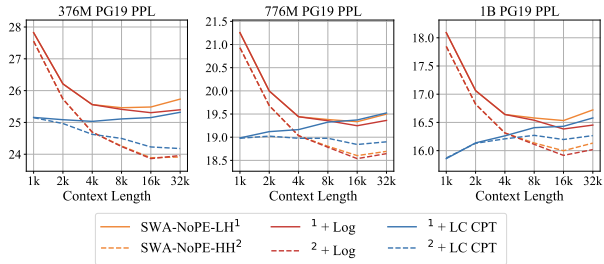
GLA and GDN behave like position-biased partners



LA heads tend toward lower hit rates and lower entropy than global NoPE heads, with differences between GLA and GDN.

Original Fig. 21(b), p. 19. 3:1 LH; rows 376M / 776M. Axes: normalized entropy and top-64 hit rate at 4K.

SWA can learn local prediction during long training



- After continual pretraining, short-context PPL improves sharply.
- Longer-context positions improve less, especially in LH models.

Original Fig. 14(a), p. 15. Solid: LH; dashed: HH. Orange: direct; red: Log; blue: continual pretraining. PPL: lower is better.

Parameter changes support a local-optimization trap

$$r_Q = \frac{\|W'_Q - W_Q\|_2}{\|W_Q\|_2}, \qquad r_K = \frac{\|W'_K - W_K\|_2}{\|W_K\|_2}.$$

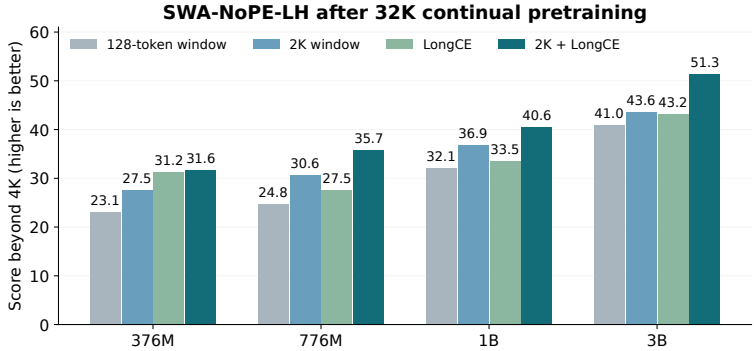
- Prime denotes parameters after long-context continued training.
- Ratios are computed per head, then averaged; HH groups are separated.
- SWA parameters change more than NoPE parameters, yet the SWA window remains 128 tokens.

Intervention suggested by the observation

Expand the branch receiving those updates; emphasize long dependencies in the training objective.

Paper, Sec. 5.1 and Fig. 15, pp. 14–16. Parameter displacement supports this interpretation; it is not a direct gradient-budget measurement.

A larger window plus LongCE improves SWA extension



Redrawn from Table 15, p. 50. SWA-NoPE-LH, 32K training. >4K averages eight RULER/BABILong cells, including 64K.

Window choice should depend on the training stage

776M SWA-NoPE-LH	128	512	2K	4K
4K training: score >4K	20.4	**22.2**	18.6	17.7
32K training: score >4K	24.8	28.4	30.6	**31.1**
32K training: All	29.4	32.6	**34.0**	33.7

- Short training: overly large windows weaken extrapolation.
- Long training: overly small windows restrict extension benefits.
- The reported 2K optimum concerns All, not every metric.

Paper, Sec. 5.2; Tables 13–14, p. 49. Short-stage rows use direct extrapolation. Extension also adjusts the RoPE base.

EME applies complementary changes to both branches

Position-biased branch

Restrict direct attention to the training-length range.

RoPE $\rightarrow$ sliding-window softmax

GLA/GDN $\rightarrow$ sliding-window LA

Global NoPE branch

Keep global access and increase logit scale.

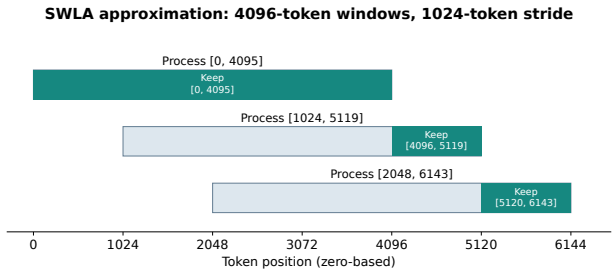
$c(L) = \log L / \log L_0$ for $L \geq L_0$

$\alpha = \text{softmax}(c(L)z)$

At $L_0 = 4096$, $L = 65536$, the logit multiplier is $4/3$.

Paper, Secs. 3.2 and 6.2, pp. 4 and 20–23. EME: Extrapolation based on the Matthew Effect. Formula expresses the stated log-base rule.

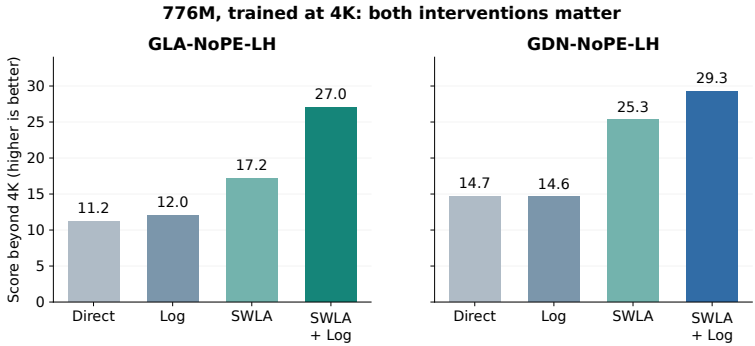
SWLA reuses LA kernels on overlapping windows



- Independent window states; concatenate only the retained outputs.
- Retained local prefixes span 3073–4096 tokens: an approximation.

Teaching diagram based on Sec. 6.2, p. 20. Window = 4096; stride = 1024. It is not an original paper figure.

Log correction becomes more useful after SWLA



Redrawn from Table 7, p. 43. 776M, 3:1 LH, trained at 4K. Same >4K aggregation for all four variants.

The 100% result is specific to SK1 at 776M

64K accuracy (%)	SK1 direct	SK1 EME	SK2 EME	SK3 EME
GLA-NoPE-LH	0	**100**	59	64
GLA-NoPE-HH	0	**100**	31	34
GDN-NoPE-LH	99	**100**	90	53
GDN-NoPE-HH	92	**100**	35	16

- All models were trained at 4K: 16× length extrapolation.
- GDN already reaches 100% on SK1 with Log alone.
- Harder SK2/SK3 cases reveal the remaining performance gaps.

Paper, Table 1, p. 23. These are RULER subtask accuracies, not overall long-context accuracy.

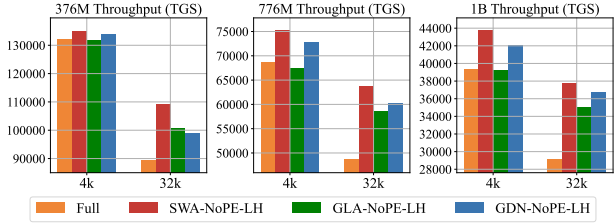
LongBench supports the trend with configuration effects

3B, LH; LongBench Avg.	4K training + Log	32K training
SWA-NoPE	**20.1**	23.7
GLA-NoPE	17.7	**25.2**
GDN-NoPE	18.1	24.9

- Input cap: 31,500 tokens; output cap: 500 tokens.
- ICL tasks give clearer evidence than difficult QA/summarization.
- At 776M LH, SWA is not initially ahead on the overall average.

Paper, Appendix A.1, pp. 32–34; Tables 18–19 and 22, pp. 53–54 and 57.

Efficiency depends on the implementation and workload



- Training throughput improves, especially at longer contexts.
- FP32 LA states and FP16 SWA-128 caches have similar sizes here.
- Overlapping SWLA repeats roughly $W/S = 4$ times the LA work; these plots do not separately benchmark that EME implementation.

Original Fig. 27(a), p. 24; Sec. 6.3. 8 H200 GPUs; TGS = tokens/GPU/s. SWLA work estimate is derived from the overlap recipe.

Three boundaries matter for using these findings

1. **Scale and task.** Up to 3B, 50B initial tokens, pretraining tasks. A 3B GLA-HH EME model scores only 5% on 64K SK1.
2. **Layer layout.** The seesaw persists in a YOCO-like ordering, but EME is less effective in that ordering.
3. **Cost.** Global NoPE retains growing KV memory; SWLA overlap adds work. Quality gains need their own system measurements.

Paper, Limitations, p. 25; Appendix A.2, pp. 34–36; Table 16, p. 51; Sec. 6.2, p. 20.

Choose the recipe for the stage and the failure mode

For training-free LA-NoPE extrapolation

Test a bounded LA range together with stronger NoPE logits. Evaluate multiple retrieval tasks and actual inference cost.

For long-context training of SWA hybrids

Revisit window size, RoPE base, and the long-dependency objective together. Inspect PPL by context length.

Mechanistic takeaway: global localization and concentrated processing are complementary roles whose balance changes with training and inference conditions.

Synthesis of Secs. 4–6. These practical interpretations retain the paper's model and protocol conditions.

Paper and discussion

Xiaoran Liu, Ziwei He, Xipeng Qiu.

Mechanics of Long-Context Hybrid Models Part 1.1: From Hybrid Attention to Hybrid Position.

arXiv:2610.10114v1, October 7, 2026.

All experimental values in this presentation come from this version. Original panels and redrawn tables are identified on their slides.

Discussion: Can a more efficient SWLA implementation retain the same gains? Can targeted branch updates test the local-learning explanation more directly?

End of main presentation: 23 slides, approximately 25 minutes. Backup slides follow.

Backup: model and optimizer configuration

	376M	776M	1B	3B
Hidden size	1024	1536	2048	3072
Intermediate size	3584	5376	7168	10752
Layers	8	12	16	24
Query / KV heads	8 / 8	12 / 12	16 / 16	24 / 24
H200 GPUs	8	8	8	16

- Vocabulary 128,256; BFloat16 training; AdamW and cosine decay.
- Initial: weight decay 0.1; warm-up 0.5B of 50B tokens.
- Continued: weight decay 0; warm-up 1B of 5B tokens.

Paper, Appendix A and Table 2, p. 32. Training throughput comparison excludes 3B to keep the GPU count fixed.

Backup: what the extended distribution measures

GLA recurrence: explanatory expansion

$$S_t = D_t S_{t-1} + k_t^\top v_t, \qquad o_t = \sum_{s=0}^{t} \left[q_t \left(\prod_{i=s+1}^{t} D_i \right) k_s^\top \right] v_s.$$

- $S_t \in \mathbb{R}^{d_k \times d_v}$; q, k, v are row vectors.
- A scalar signed coefficient a gives $\|aI\|_2 = |a|$.
- Example: coefficients $(+2, -2)$ normalize to $(0.5, 0.5)$; sign cancellation disappears.
- Normalized entropy also depends on the denominator $\log L$.

Explanatory derivation from the GLA gate product in Sec. 6.2, p. 21, and Eq. (1), p. 18. Interpret as value-path sensitivity.

Backup: the reported averages weight length cells

Reported column	Included scores	Number of cells
RULER	4K, 8K, 16K, 32K, 64K	5
BABILong	0K, 2K, 4K, 8K, 16K, 32K, 64K	7
$\leq$4K	RULER 4K; BABILong 0K/2K/4K	4
$>$4K	Both benchmarks at 8K/16K/32K/64K	8
All	All listed benchmark–length scores	12

$$\text{All} \approx \frac{5\,\text{RULER Avg.} + 7\,\text{BABILong Avg.}}{12}.$$

The approximation sign accounts for displayed rounding. These are not two equally weighted benchmark averages.

Paper, table headers and entries in Tables 7–15, pp. 43–50.

Backup: SWLA and Log solve different parts of retrieval

776M GLA-NoPE-LH; 64K accuracy (%)	SK1	SK2	SK3
Direct	0	0	0
Log	0	0	0
SWLA	97	0	3
SWLA + Log	**100**	**59**	**64**

SWLA can rescue the easiest subtask while the harder subtasks still need effective global aggregation.

At 32K SK1, the same EME model scores 96%: 100% at 64K does not mean 100% at every tested length.

Paper, Table 1, p. 23. Trained at 4K; no additional training for these interventions.

Backup: the headline does not hold at every scale

GLA-NoPE-HH + EME; 64K accuracy (%)	SK1	SK2	SK3
376M	100	49	37
776M	100	31	34
1B	64	68	66
3B	5	100	73

The intervention is useful across settings, but extrapolation quality depends strongly on the task and model. Larger scale does not guarantee a higher SK1 score here.

Paper, Table 1, p. 23; Table 16, p. 51. All models initially trained at 4K on 50B tokens.

Backup: additional checks and unresolved choices

776M, YOCO-like LH; All	SWA-NoPE	GLA-NoPE	GDN-NoPE
Short training + Log	30.6	21.7	25.6
Long continued training	30.6	36.1	31.4

- YOCO-like here means placing efficient layers at the bottom.
- Four learning rates retain the main seesaw trend.
- Smaller SWA RoPE bases and dropping RoPE during continued training show no consistent optimum across configurations.

Paper, Appendix A.2–A.3, pp. 34–37; Table 24, p. 59. EME is less effective in the YOCO-like ordering.