MECHANICS OF LONG-CONTEXT HYBRID MODELS 1.1: FROM HYBRID ATTENTION TO HYBRID POSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

The architectural design of Large Language Models (LLMs) is shifting from traditional full-attention-only models to hybrid models, which combine different attention modules to improve long-context efficiency and performance in length extrapolation and context extension. To explain why hybrid models work and how to design them better, we propose **Mechanics of Long-Context Hybrid Models**. As Part 1.1 of this series, we begin with hybrids of full attention and either sliding-window attention (SWA) or gated variants of linear attention (LA), represented by GLA and GDN. We first observe a **Seesaw Effect in Context Extension**: SWA hybrids perform better in length extrapolation, while LA hybrids perform better after context extension. We interpret this from the perspective of hybrid position. We find that SWA hybrids suffer from a **Short-Context Learning Trap**, **Short-Window Weariness**, and **Long-Window Laziness**, requiring extended windows to enhance performance in continual long-context pretraining. In contrast, we summarize the **Matthew Effect of Hybrid Position Extrapolation** and propose **Sliding-Window Linear Attention** for LA hybrids, achieving up to 16× training-free length extrapolation with 100% retrieval accuracy on NIAH-SK1.

1 INTRODUCTION

The architectural design of Large language models (LLMs) (Brown et al., 2020; Sun et al., 2024a; Cai et al., 2024; Dubey et al., 2024; Guo et al., 2025) is shifting from traditional full-softmax-attention-only models (Vaswani et al., 2017) to hybrid models (Jamba et al., 2024; Agarwal et al., 2025; Gemma et al., 2026; Li et al., 2025; Kimi et al., 2026; Qwen, 2026b; MiMo, 2026; Zeng et al., 2026; Xu et al., 2026), which combine different attention modules across layers or heads and introduce different position biases to improve long-context efficiency and performance under context extension and length extrapolation (Wang et al., 2025; Kimi et al., 2025; Chen et al., 2026; Merrill et al., 2026). To answer **why hybrid architectures work** and **how to design them better** for long-context LLMs, we propose **Mechanics of Long-Context Hybrid Models**, through systematic training, evaluation, mechanism analysis, and closed-loop validation.

- **Why Hybrid?** Hybrid models have become an important paradigm in LLM design (Agarwal et al., 2025; Kimi et al., 2026; Qwen, 2026b). Compared with traditional RoPE-based full-attention-only models (Su et al., 2024; Touvron et al., 2023), hybrid models not only offer higher computational and memory efficiency (Liu et al., 2024a; 2025a; Xu et al., 2026), but also provide better performance in retrieval, state tracking, and length extrapolation (Puvvada et al., 2025; Hu et al., 2026a;c; Kimi et al., 2025; Merrill et al., 2026; Yang et al., 2026).
- **Why Long-Context?** Long context is one of the most important application scenarios for current hybrid models (Puvvada et al., 2025; Kimi et al., 2025; Merrill et al., 2026). Meanwhile, million-token-level contexts in the agentic era impose strong constraints on model architectures (Li et al., 2025; Kimi et al., 2026; Xu et al., 2026). Relying solely on full attention is unrealistic, while any single efficient attention is unreliable (Yang et al., 2024a; 2025b; Kimi et al., 2025). This makes hybrid models one of the most worth-studying fields in long-context LLMs (Wang et al., 2025; Merrill et al., 2026; Qiao et al., 2026).
- **Why Mechanics?** There has already been substantial work discussing hybrid models from the perspectives of downstream task evaluation, scaling curves, or system efficiency (Puvvada et al., 2025; Wang et al., 2025; Merrill et al., 2026; Qiao et al., 2026;

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.

Tan et al., 2026), but these results often remain at the level of comparing the phenomenon of "which model is better" (Merrill et al., 2026; Qiao et al., 2026; Jolicoeur-Martineau et al., 2026). We instead focus on observing and explaining the differences in long-context performance between hybrid models and traditional models—as well as among different hybrid models—thereby gaining insight into the mechanisms underlying length extrapolation and context window expansion in hybrid models.

On account of this long journey involving comparisons across multiple hybrid modules, hybrid paradigms, and training scenarios, this paper is only Part 1.1 of the whole series.

- **In Part 1**, we first discuss the hybrid of memory-constant sliding-window attention (SWA) or linear attention (LA) with full attention, one of the most widely used hybrid models (Agarwal et al., 2025; Gemma et al., 2026; MiMo, 2026; Qwen, 2026b; Kimi et al., 2026).
- **In Part 1.1**, we analyze the long-context language modeling (Rae et al., 2019; Fang et al., 2024), retrieval (Kamradt, 2023; Hsieh et al., 2024), and state-tracking abilities (Kuratov et al., 2024) in pretraining stage, focusing on the length extrapolation (i.e. training-free generalization beyond the pretraining context length) and context extension (i.e. long-context continual pretraining that expands context length), comparing RoPE-based full-attention models, RoPE-NoPE hybrid-position full-attention models, SWA hybrid models, and LA hybrid models represented by GLA/GDN (Yang et al., 2024a; 2025b; Qwen, 2026b)

Specifically, the contributions of this paper can be summarized as follows:

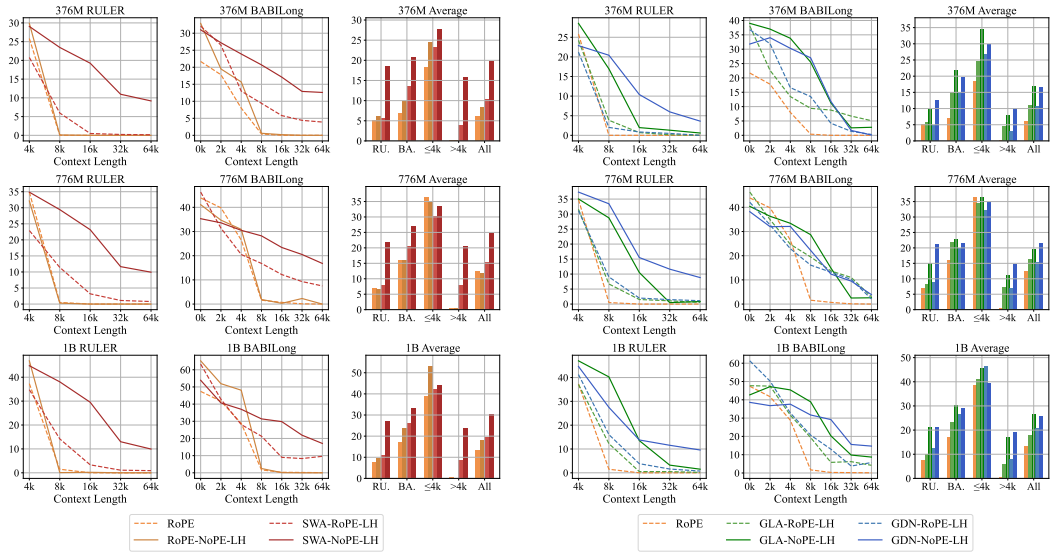
- We analyze the long-context performance of hybrid models from the perspective of **hybrid position**, through extensive experiments on RoPE-based full attention, RoPE-NoPE hybrid, SWA hybrid, and GLA/GDN hybrid, both layer-wise and head-wise, across model scales of 376M, 776M, 1B and 3B, under length extrapolation and context extension, and find the **Seesaw Effect** of context extension in hybrid models.
- **On SWA hybrids**, we demonstrate that although SWA hybrids exhibit clear advantages in length extrapolation, they suffer from the **Short-Context Learning Trap** in context extension. To address this, we improve SWA hybrids by enlarging the window size and applying the LongCE loss, and further summarize the underlying **Short-Window Weariness** and **Long-Window Laziness**.
- **On LA hybrids**, we find that although LA hybrids perform better in long-context continual pretraining, they are weaker in length extrapolation. To address this, we propose **Sliding-Window Linear Attention**, achieving 16× length extrapolation with 100% retrieval accuracy, and summarize the underlying **Matthew Effect** in hybrid position extrapolation.

2 OBSERVATION

2.1 SETUP

We conduct 4k short-context pretraining with 50B tokens and 32k long-context continual pretraining with 5B tokens at 376M, 776M, and 1B model sizes, and also verify important conclusions by pretraining on a 3B model size with the configuration detailed in Table 2 in Appendix A. We abbreviate layer-wise hybrid as LH and head-wise hybrid as HH, and use RoPE and NoPE to denote full-attention using the corresponding position embedding. For example, SWA-NoPE-LH denotes a model where SWA and NoPE full attention are mixed by layer with a ratio of 3:1 (Wang et al., 2025; Qwen, 2026b; Puvvada et al., 2025). We set the window size of SWA to 128, following GPT-OSS (Agarwal et al., 2025) and Mimo-v2.5 (MiMo, 2026), and set the rotary base of SWA and RoPE full attention to 10000 (Su et al., 2024), while GLA and GDN follow their original papers (Yang et al., 2024a; 2025b; Yang & Zhang, 2024). We only use MHA, since GDN does not support GQA.

We use the PG19 dataset (Rae et al., 2019) for perplexity (PPL) and longPPL (Fang et al., 2024) evaluation, and use RULER (Hsieh et al., 2024) and BABILong (Kuratov et al., 2024) to evaluate retrieval and state-tracking abilities, respectively. In addition, we also report the average scores within 4k and beyond 4k as references for in-domain and out-of-domain performance. We report the performance of the layer-wise hybrids on RULER and BABILONG in the main text. Training details, perplexity evaluations, and the performance of the head-wise hybrids are provided in Appendix A.



(a) Results on Full Attention and SWA Hybrid.

(b) Results on Full Attention and GLA/GDN Hybrid.

Figure 1: Comparison of long-context performance under RULER and BABILong benchmarks among different layer-wise hybrid models after 4k context length pretraining.

2.2 Short-Context Training

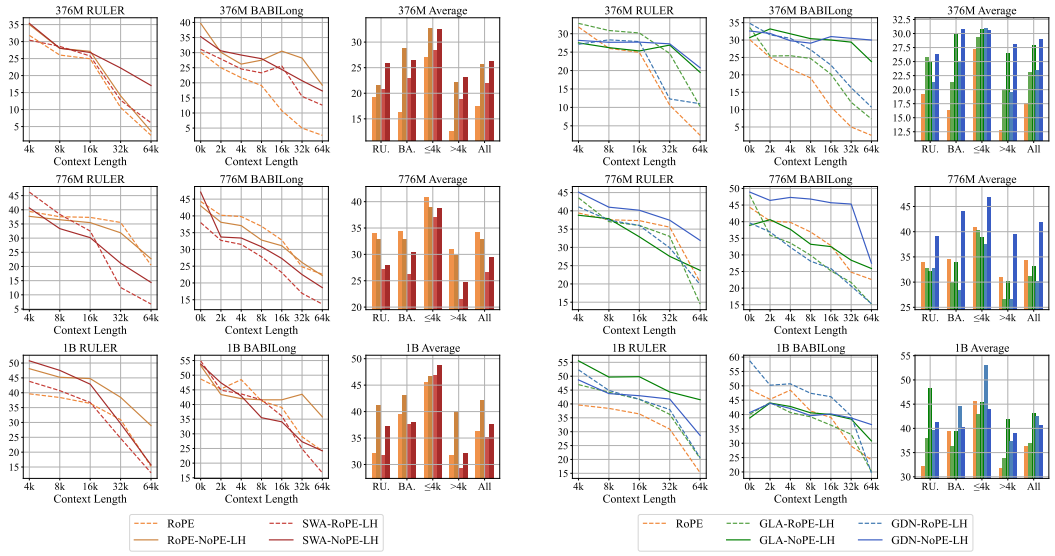
Figure 1 and Figure 11 show the results of layer-wise hybrids, while Figure 13 and Figure 12 show those of head-wise hybrids. Compared with the RoPE-only full attention model, the hybrid models exhibit better performance within the training length, even for RoPE-NoPE hybrid full attention models (Yang et al., 2026). In SWA hybrids and LA hybrids, removing the positional embedding in the full attention layer still enhances long-context performance and length extrapolation (Puvvada et al., 2025; Jamba et al., 2024), in both layer-wise and head-wise hybrids.

We also verify the effectiveness of RoPE-NoPE extrapolation in an SWA-NoPE hybrid manner (Yang et al., 2026; Puvvada et al., 2025) rather than positional interpolation (bloc97, 2023b;a; Chen et al., 2023; Peng et al., 2024). For RoPE attention, we retain only the attention window of the nearest pretraining context length, 4k. For NoPE attention, we use a log-scaled attention extrapolation, increasing the attention temperature to prevent attention entropy from increasing as the context grows, with the log base equal to the pretraining context length (Su, 2023; Liu et al., 2024b). As shown in Figure 16 in Appendix A, this hybrid extrapolation method significantly outperforms RoPE or RoPE-NoPE hybrids using Dynamic NTK interpolation (bloc97, 2023a; Liu et al., 2024b).

Based on the above observation, we suppose that hybrid position bias is key to achieving better performance on long contexts in the hybrid models. **Hybrid models improve long-context performance and length extrapolation through a hybrid position of NoPE and other position biases**, not only in RoPE-NoPE hybrids and SWA-NoPE hybrids, but also in GLA/GDN-NoPE hybrids, since the gate mechanism of linear attention can also be regarded as a data-dependent implicit positional bias (Lin et al., 2025; Zhong et al., 2025; Yang et al., 2025c).

2.3 Long-Context Training

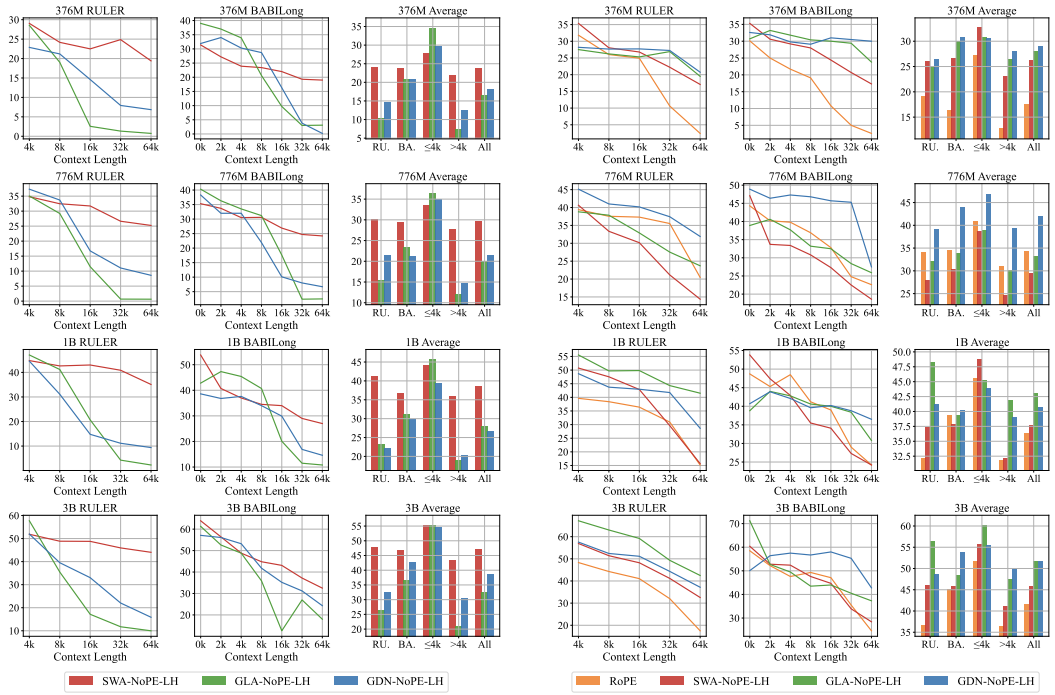
We further conduct long-context continual pretraining. Figure 2 shows results on layer-wise hybrids, while Figure 14 shows those on head-wise hybrids. Notably, there exists a very interesting phenomenon. SWA hybrids, which perform best in length extrapolation, underperform the LA hybrids after context extension, as shown in Figure 3 and Figure 15, and may even fail to surpass their performance under log-scaled NoPE extrapolation, which is more remarkable in layer-wise hybrids. In contrast, LA hybrids benefit more from context extension during long-context continual pretraining, but lag behind SWA hybrids in length extrapolation. We call this reversal the **Seesaw**



(a) Results on Full Attention and SWA Hybrid.

(b) Results on Full Attention and GLA/GDN Hybrid.

Figure 2: Comparison of long-context performance under RULER and BABILong benchmarks among different layer-wise hybrid models after 32k context length continual pretraining.



(a) Performance of log-scale NoPE extrapolation.

(b) Performance after long-context pretraining.

Figure 3: Comparison of the long-context performance of the SWA/GLA/GDN-NoPE layer-wise hybrid models under log-scale NoPE extrapolation after short-context pretraining (left) and after long-context continual pretraining (right), exhibiting a remarkable seesaw effect.

Effect of Context Extension in Hybrid Models: **SWA-NoPE hybrids are stronger for length extrapolation after short-context pretraining, whereas LA-NoPE hybrids benefit more from long-context continual pretraining, especially in layer-wise mixing**. In the following sections,

we will analyze the reason why such phenomenon exists and improve the context extension of SWA hybrids and the length extrapolation of LA hybrids.

3 Discussion

3.1 Short-Context Learning Trap

We start the discussion by observing the perplexity curves. Figure 4 compares the perplexity curves of SWA-NoPE, RoPE-NoPE, GLA-NoPE and GDN-NoPE hybrids before and after long-context continual pretraining. For most hybrids apart from SWA-NoPE hybrids, the perplexity curve before long-context training rises clearly beyond the pretraining context length. After long-context continual pretraining, the curve becomes much flatter or steadily decreases, indicating that the model can support longer contexts and use context to predict the next token better.

However, SWA-NoPE hybrids show the opposite behavior. Because they already have strong extrapolation ability, their perplexity curves keep decreasing within 32k before long-context continual pretraining. After long-context continual pretraining, especially for layer-wise SWA-NoPE hybrids, the perplexity curve instead shows an upward trend: **the loss on long-context positions decreases only slightly**, while **the loss on short-context positions drops abnormally strongly**, and this

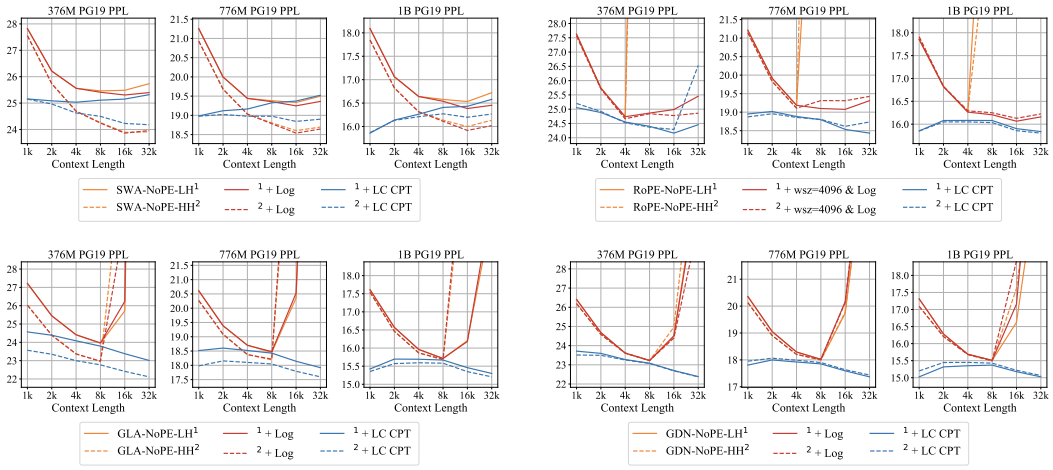
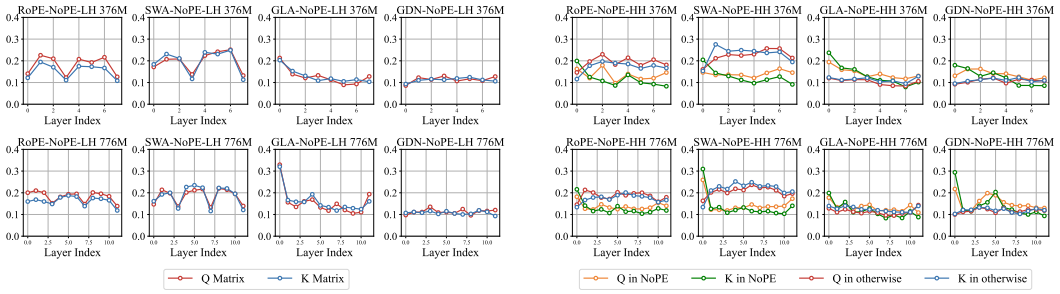


Figure 4: Perplexity curves of different hybrid models under direct extrapolation, log-scale NoPE extrapolation (Log), and long-context continual pretraining (LC-CPT). The perplexity curve of SWA-NoPE-LH shows a remarkable increasing trend as the context length grows.



(a) Comparison of layer-wise hybrid models.

(b) Comparison of head-wise hybrid models.

Figure 5: Comparison of the change rates of the $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$ matrices in attention modules across different hybrid models before and after long-context training. In layer-wise hybrids, 1 NoPE layer is inserted every 3 RoPE/SWA/GLA/GDN layers. In head-wise hybrids, each layer contains the $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$ matrices of NoPE as well as those of the other RoPE/SWA/GLA/GDN modules.

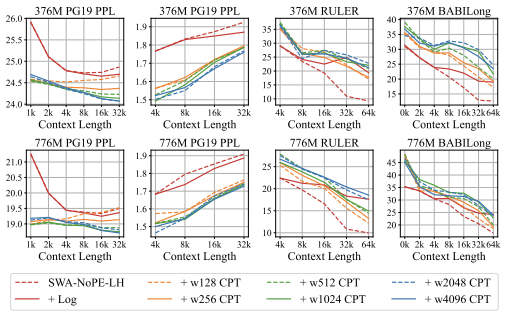


Figure 6: Long-context performance of SWA-NoPE-LH with different window sizes (wsz) in long-context continual pretraining (CPT).

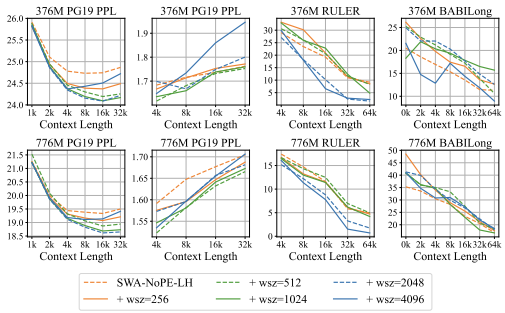


Figure 7: Long-context performance of SWA-NoPE-LH with different window sizes (wsz) in short-context pretraining.

phenomenon becomes more obvious as the model size increases. This suggests that the additional long-context training signal is not effectively converted into improved long-context predictions. Long-context continual pretraining of SWA-NoPE hybrids focuses too much on short contexts, and as a result, LA-NoPE hybrids surpass SWA-NoPE hybrids after long-context continual pretraining.

To further examine this hypothesis of short-context learning traps, we plot the parameter change ratios in Figure 5, comparing SWA-NoPE, RoPE-NoPE, GLA-NoPE, and GDN-NoPE hybrids. The x-axis represents different layers, and the y-axis represents the parameter change ratios of the $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$ matrices. Taking $\boldsymbol{W}_Q$ as an example, let $\boldsymbol{W}_Q$ and $\boldsymbol{W}'_Q$ denote its values before and after long-context continual pretraining. We define the change ratio as $\|\boldsymbol{W}'_Q - \boldsymbol{W}_Q\|_2 / \|\boldsymbol{W}_Q\|_2$. For head-wise hybrids, we separately aggregate NoPE heads and non-NoPE heads in each layer. For layer-wise hybrids, we do not make this distinction, and the non-NoPE layers and NoPE layers are interleaved at a 3:1 ratio.

We find that for GLA-NoPE and GDN-NoPE hybrids, the parameter change ratios of GLA/GDN and NoPE attention are close. In contrast, for SWA and RoPE attention, especially SWA, the parameter change ratio is significantly larger than that of NoPE attention in both layer-wise and head-wise hybrids. This suggests that because they are explicitly position-biased modules, **SWA and RoPE attention undergo relatively larger optimization during long-context continual pretraining**, in SWA-NoPE and RoPE-NoPE hybrids. For SWA, however, the window size remains fixed, so the range of context dependencies that it can model is still limited. Therefore, its abnormal focus on this bounded local branch is a possible contributor to the short-context learning trap.

3.2 Context Extension of SWA Hybrid

Motivated by Figure 5, we enlarge the SWA window size during long-context continual pretraining of layer-wise SWA-NoPE hybrids. The idea is to use the abnormal focus on the SWA branch to capture longer context dependencies. Specifically, we increase the window size from 128 to 256, 512, 1024, 2048, and 4096. Following the scaling law of RoPE-based extrapolation (Liu et al., 2024b), we also compute the rotary base corresponding to each enlarged window size when the initial window size is 128, and the original rotary base is 10000. Figure 6 and Figure 8 compare the PG19 perplexity, RULER, and BABILong performance of SWA hybrids after enlarging the window. The results show that the enlarged-window configurations improve downstream long-context performance, and the 2048-window configuration gives the best results among the evaluated settings.

In addition, enlarging the window only exploits the short-context learning trap rather than emphasizing long-context dependencies directly. Therefore, we also use LongCE (Fang et al., 2024) as the training objective for layer-wise SWA-NoPE hybrids in long-context continual pretraining. As shown in Figure 10, both enlarging the window to 2048 and using LongCE improve the long-context performance of SWA-NoPE layer-wise hybrids, and the two methods can be combined. With both methods, SWA-NoPE-LH becomes competitive with GDN-NoPE-LH across different model sizes. Since the short-context trap in SWA hybrid long-context continual pretraining may not be limited to SWA-NoPE layer-wise hybrids, we further apply the same two strategies to SWA-RoPE layer-

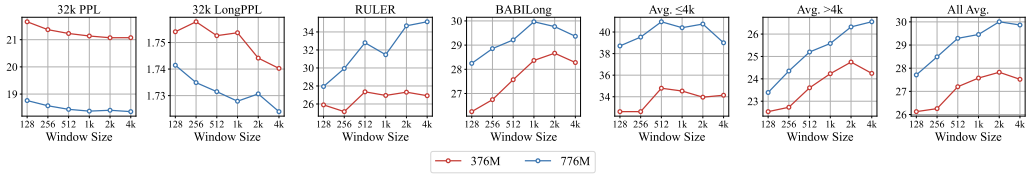


Figure 8: Long-context performance as SWA-NoPE-LH's window size increases in long-context pretraining. Larger window sizes show better context extension, implying short-window weariness.

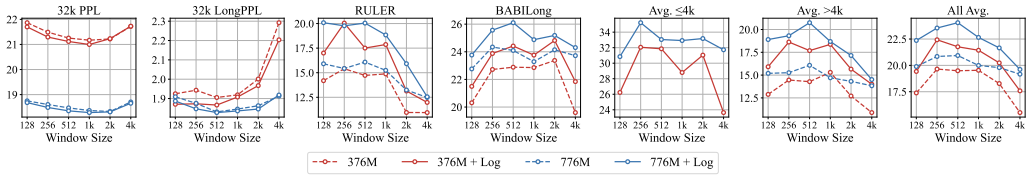
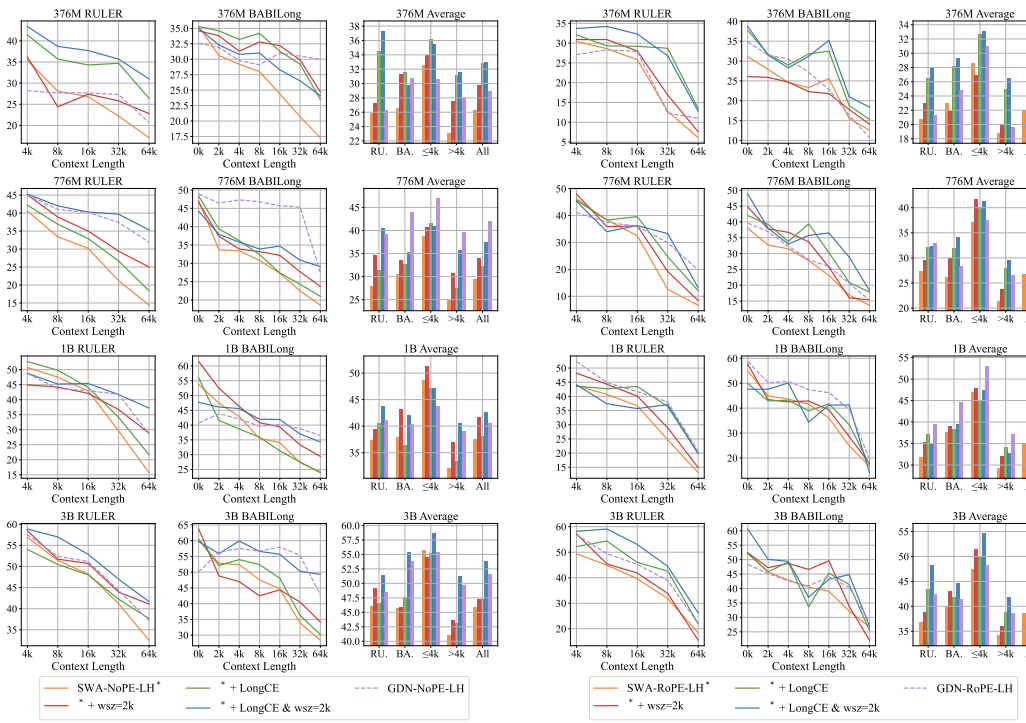


Figure 9: Long-context performance as SWA-NoPE-LH's window size increases in short-context pretraining. Shorter window sizes show better length extrapolation, implying long-window laziness.



(a) Results of SWA-NoPE-LH.

(b) Results of SWA-RoPE-LH.

Figure 10: Long-context performance of SWA-NoPE-LH in context extension enhanced by LongCE and enlarged window size. Both weaken the seesaw effect and also contribute to SWA-RoPE-LH.

wise hybrids. Enlarging the window and using LongCE also improves the long-context training performance of SWA-RoPE-LH at different scales, making it close to GDN-RoPE-LH.

These results show that keeping a small SWA window during long-context continual pretraining harms context extension. **SWA hybrids suffer from the short-context learning trap and require enlarged windows as well as training strategies that emphasize long-context learning**. We call this phenomenon **Short-Window Weariness of SWA Hybrid**. At the same time, some previous works have reported that for SWA hybrids, shorter windows can be more advantageous than longer windows, because they better activate the potential of NoPE full-attention layers (Hu et al., 2026c;

Qiao et al., 2026); this phenomenon is sometimes called **Long-Window Laziness**. We argue that the two observations are not contradictory. In fact, we can also verify this in Figure 7 and Figure 9.

The coexistence of short-window weariness and long-window laziness indicates that **SWA hybrids may require fine-grained adjustment of positional bias across short-context and long-context training stages,** just as RoPE full-attention models need to tune the rotary base and scaling factor in context extension (Chen et al., 2023; Peng et al., 2024; Liu et al., 2024b; Xiong et al., 2024). We summarize the coexistence of short-window weariness and long-window laziness as follows: **in short-context pretraining, SWA hybrids with shorter window show better extrapolation**, while **in long-context pretraining, SWA hybrids with larger window show better context extension**.

3.3 Length Extrapolation of LA Hybrid

In RoPE-NoPE hybrids, we achieve length extrapolation beyond traditional NTK interpolation by treating RoPE attention as a sliding-window attention module with a 4k window, while keeping NoPE attention globally aware and strengthening it in a log-scale correction (Yang et al., 2026; Puvvada et al., 2025) as shown in Figure 16. For LA-NoPE hybrids, since the gate mechanism of linear attention can also be regarded as a data-dependent position embedding (Lin et al., 2025; Zhong et al., 2025; Yang et al., 2025c), we analogously propose **Sliding-Window Linear Attention** (SWLA), which applies a sliding window to linear attention with gates. This may sound counterintuitive, because it appears to conflict with the recurrent nature of linear attention, but it can still be implemented. Without modifying FLA kernels (Yang & Zhang, 2024), we use an approximation similar to sliding-window perplexity evaluation (Press et al., 2022). Suppose the window size is 4096 and the stride is 1024. We can use the original GLA or GDN operator to compute attention on chunks $[0, 4095], [1024, 5119], [2048, 6143], \cdots$ in parallel, and then extract the outputs at $[0, 4095], [4096, 5119], [5120, 6143], \cdots$ before stacking them back together. This implementation is independent of the internal details of a specific LA variant.

From the internal logic of linear attention, using GLA as an example, the original interaction between $\boldsymbol{q}_t$ and $\boldsymbol{k}_s$ is multiplied by $\prod_{i=s+1}^{t} \mathrm{Diag}(\boldsymbol{\alpha}_i)$. After applying Sliding-Window Linear Attention, the number of accumulated terms in this product is upper-bounded, implying that the relative-position signal is also restricted to the training length. As shown in Figures 17 and 18, GLA-NoPE and GDN-NoPE hybrids with Sliding-Window Linear Attention achieve clear improvements in length extrapolation. We also highlight the results on the NIAH-SK1, SK2, and SK3 tasks in RULER, which are the most frequently reported tasks in hybrid model research. GLA-NoPE and GDN-NoPE

	NIAH-SK1					**NIAH-SK2**					**NIAH-SK3**				
	4k	8k	16k	32k	64k	4k	8k	16k	32k	64k	4k	8k	16k	32k	64k
776M Short															
GLA-NoPE-LH	99.0	95.0	80.0	0.0	0.0	99.0	69.0	0.0	0.0	0.0	90.0	88.0	0.0	0.0	0.0
+ Log	99.0	99.0	98.0	0.0	0.0	99.0	68.0	1.0	0.0	0.0	90.0	91.0	0.0	0.0	0.0
+ SWLA	99.0	**100.0**	**100.0**	97.0	97.0	99.0	74.0	5.0	0.0	0.0	90.0	56.0	21.0	8.0	3.0
+ SWLA & Log	99.0	**100.0**	**100.0**	96.0	**100.0**	99.0	92.0	87.0	72.0	59.0	90.0	**79.0**	**75.0**	**76.0**	**64.0**
GLA-NoPE-HH	**100.0**	**100.0**	68.0	82.0	0.0	**100.0**	71.0	0.0	0.0	0.0	65.0	45.0	0.0	0.0	0.0
+ Log	**100.0**	**100.0**	98.0	95.0	72.0	**100.0**	68.0	0.0	0.0	0.0	65.0	49.0	0.0	0.0	0.0
+ SWLA	**100.0**	**100.0**	**100.0**	92.0	75.0	**100.0**	**100.0**	82.0	32.0	1.0	65.0	62.0	44.0	9.0	1.0
+ SWLA & Log	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	97.0	66.0	31.0	65.0	68.0	65.0	62.0	34.0
GDN-NoPE-LH	**100.0**	**100.0**	**100.0**	**100.0**	99.0	99.0	98.0	2.0	0.0	0.0	62.0	71.0	5.0	0.0	0.0
+ Log	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	99.0	98.0	9.0	1.0	0.0	62.0	70.0	9.0	0.0	0.0
+ SWLA	**100.0**	**100.0**	**100.0**	**100.0**	99.0	99.0	97.0	96.0	92.0	53.0	62.0	64.0	61.0	56.0	31.0
+ SWLA & Log	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	99.0	97.0	**99.0**	**99.0**	**90.0**	62.0	67.0	64.0	66.0	53.0
GDN-NoPE-HH	**100.0**	**100.0**	**100.0**	99.0	92.0	**100.0**	95.0	2.0	0.0	0.0	**94.0**	38.0	0.0	0.0	0.0
+ Log	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	93.0	3.0	0.0	0.0	**94.0**	43.0	0.0	0.0	0.0
+ SWLA	**100.0**	99.0	**100.0**	95.0	92.0	**100.0**	99.0	64.0	4.0	0.0	**94.0**	68.0	31.0	5.0	2.0
+ SWLA & Log	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	97.0	72.0	35.0	**94.0**	71.0	65.0	47.0	16.0

Table 1: Comparison of the 776M GLA/GDN-NoPE hybrids on the NIAH-SK1, SK2, and SK3 tasks. Applying our SWLA and a log scale for NoPE attention achieves the best length extrapolation.

with SWLA and log-scaled NoPE attention achieve 16× training-free length extrapolation while maintaining almost 100% retrieval accuracy on SK1, as shown in Table 1 and Table 16.

Furthermore, we find that directly adding log-scale NoPE attention in LA-NoPE hybrids brings only limited improvement, but after applying SWLA, log-scale correction becomes much more effective. This also implies that even for linear attention with implicit positional bias, extrapolation should still restrict the relative-position range. This means **in length extrapolation of hybrid-position models, position-biased attention should use a sliding window to restrict its relative-position information, while NoPE attention should strengthen its global perception**. We call this phenomenon the **Matthew Effect of Length Extrapolation in Hybrid Position**. In fact, the effectiveness of SWLA is also consistent with reports from other work. For example, in Figure 5 of Chen et al. (2026), the best extrapolating linear-attention hybrid is HypeNet-Lightning, and long-range decay is a key feature of Lightning Attention (Qin et al., 2024). **Our approach based on SWLA further amplifies this property through the sliding window, breaks the traditional boundaries between SWA and LA, and is in principle compatible with arbitrary LA variants.**

In addition to the aforementioned discussion of length extrapolation and context extension phenomena in hybrid models, we also include further validation experiments and open discussions in Appendix A. In Appendix A.2, we first demonstrate that the Seesaw Effect persists under different layer-wise hybrid layout (Sun et al., 2024b) and long-context learning rate. In Appendix A.3, we include discussions on topics where conclusions are not yet definitive, such as transfer training schemes, the selection of the rotary base for SWA in SWA hybrids, and performance differences within LA-based hybrids. In Appendix A.4, we also add a brief efficiency comparison. Furthermore, Appendix B provides detailed evaluation scores, along with results from additional standard short-context tasks (HuggingFace, 2023). We also add Limitations section in Appendix A.6 and yet believe our work contributes to a better understanding of long-context phenomena in hybrid architectures.

4 Related Work

LLM architecture design has shifted from traditional full-attention-only models to hybrid models (Jamba et al., 2024; Agarwal et al., 2025; Gemma et al., 2026; Li et al., 2025; Kimi et al., 2026; Qwen, 2026b; MiMo, 2026; Zeng et al., 2026; Xu et al., 2026). For example, Qwen-3.5 (Qwen, 2026b;a), Kimi-K3 (Kimi et al., 2025; 2026), and OLMo-Hybrid (Merrill et al., 2026) mix full softmax attention (Ainslie et al., 2023; Liu et al., 2024a) or sparse attention (Liu et al., 2025a) with linear attention (Yang et al., 2025b; Kimi et al., 2025). Gemma4 (Gemma et al., 2026), Inkling (Thinking Machines, 2026), and Mimo-v2 (MiMo, 2026) mix full softmax attention with sliding-window softmax attention. GLM-5.3 (Zeng et al., 2026; Bai et al., 2026) and MiniMax-M3 (Minimax, 2026; Lai et al., 2026) adopt sparse-attention hybrids (Liu et al., 2025a; Gao et al., 2026; Hu et al., 2026b), while DeepSeek-v4 (Xu et al., 2026) uses compressed-attention hybrids (Chu et al., 2026).

These hybrid models can be more computationally and memory efficient than traditional models, while achieving downstream performance close to or even better than traditional full-attention models. Among them, hybrids with fixed memory complexity, especially linear attention (LA) and sliding-window attention (SWA) hybrids, after the early exploration (De et al., 2024; Botev et al., 2024; Ren et al., 2025), have since been widely adopted (Jamba et al., 2024; Agarwal et al., 2025; Li et al., 2025; Kimi et al., 2026). They have also inspired fine-tuning attempts (Lan et al., 2025; Chen et al., 2026; Lan et al., 2026), refined efficiency optimizations (Sun et al., 2024b; Hu et al., 2026a) and long-context analysis detailed in Appendix A.5. Therefore, we start from this class of hybrids to study why hybrid models are effective and how to design them more effectively.

5 Conclusion

In this paper, we conduct extensive experiments to validate the effectiveness of the hybrid model in length extrapolation and context extension from the perspective of hybrid position. We first discover the Seesaw Effect in context extension of the SWA hybrid models versus LA hybrid models. For SWA hybrids, we enlarge the window during extended training to overcome its short-context learning trap. For LA hybrids, we propose the Sliding-Window Linear Attention, achieving 16× training-free length extrapolation while maintaining 100% retrieval accuracy and summarize the Matthew Effect in extrapolation for hybrid models with hybrid position.

AI USE STATEMENT

We used generative AI tools only for language-related assistance, including checking grammar, style, and clarity. We did not use generative AI tools to generate research ideas, formulate hypotheses, design experiments, implement methods, analyze or interpret results, or produce scientific claims. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for the final content of this paper.

ETHICS STATEMENT

This study does not involve human subjects, direct interaction with individuals, or the collection of new personal data. We use publicly available text corpora and follow the applicable dataset terms and licenses. To the best of our knowledge, the experiments do not target high-risk applications or introduce additional privacy, safety, or discrimination risks beyond those inherent in the underlying pretrained data. The authors take responsibility for the integrity and accurate reporting of the results.

REPRODUCIBILITY STATEMENT

We provide the model configurations and training details in Appendix A, including the model sizes, context lengths, data sources, optimization settings, training budgets, and hardware. The evaluation results are detailed in Appendix B. To ensure the reproducibility of and to support the open-source community, we plan to release the code and trained checkpoints after publication, and expect these as a reference for future work on long-context hybrid models, facilitating progress in this field.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b and gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

Yushi Bai, Qian Dong, Ting Jiang, Xin Lv, Zhengxiao Du, Aohan Zeng, Jie Tang, and Juanzi Li. Indexcache: Accelerating sparse attention via cross-layer index reuse. *arXiv preprint arXiv:2603.12201*, 2026.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7432–7439. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6239. URL https://doi.org/10.1609/aaai.v34i05.6239.

bloc97. Dynamically scaled rope further increases performance of long context llama with zero fine-tuning, July 2023a. URL https://www.reddit.com/r/LocalLLaMA/comments/14mrgpr/dynamically_scaled_rope_further_increases/.

bloc97. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation., June 2023b. URL https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.

Aleksandar Botev, Soham De, Samuel L Smith, Anushan Fernando, George-Cristian Muraru, Ruba Haroun, Leonard Berrada, Razvan Pascanu, Pier Giuseppe Sessa, Robert Dadashi, et al. Recurrentgemma: Moving past transformers for efficient open language models. *arXiv preprint arXiv:2404.07839*, 2024.

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

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Yingfa Chen, Zhen Leng Thai, Zihan Zhou, Zhu Zhang, Xingyu Shen, Shuo Wang, Chaojun Xiao, Xu Han, and Zhiyuan Liu. Hybrid linear attention done right: Efficient distillation and effective architectures for extremely long contexts. *arXiv preprint arXiv:2601.22156*, 2026.

Chenglong Chu, Guorui Zhou, Guowang Zhang, Han Li, Hao Peng, Hongtao Cheng, Hui Wang, Jian Liang, Jiangxia Cao, Kun Gai, et al. Kwai summary attention technical report. *arXiv preprint arXiv:2604.24432*, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018. URL `http://arxiv.org/abs/1803.05457`.

John Cooper, Ilias Diakonikolas, Mingchen Ma, and Frederic Sala. Expressivity-efficiency tradeoffs for hybrid sequence models. *arXiv preprint arXiv:2603.08859*, 2026.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, et al. Hymba: A hybrid-head architecture for small language models. *arXiv preprint arXiv:2411.13676*, 2024.

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

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling? *arXiv preprint arXiv:2410.23771*, 2024.

Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

Yizhao Gao, Jianyu Wei, Qihao Zhang, Yu Cheng, Shimao Chen, Zhengju Tang, Zihan Jiang, Yifan Song, Hailin Zhang, Liang Zhao, et al. Hysparse: a hybrid sparse attention architecture with oracle token selection and kv cache sharing. *arXiv preprint arXiv:2602.03560*, 2026.

Yoav Gelberg, Koshi Eguchi, Takuya Akiba, and Edoardo Cetin. Extending the context of pre-trained llms by dropping their positional embedding. In *International Conference on Learning Representations*, volume 2026, pp. 60170–60197, 2026.

Gemma, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Xiang Hu, Jiaqi Leng, Jun Zhao, Kewei Tu, and Wei Wu. Hardware-aligned hierarchical sparse attention for efficient long-term memory access. *Advances in Neural Information Processing Systems*, 38:88925–88950, 2026a.

Xiang Hu, Xinyu Wei, Hao Gu, Minshen Zhang, Tian Liang, Huayang Li, Lei Zhu, Yan Wang, Sirui Han, Yushi Bai, et al. Hierarchical sparse attention done right: Toward infinite context modeling. *arXiv preprint arXiv:2607.02980*, 2026b.

Xiang Hu, Zhanchao Zhou, Ruiqi Liang, Zehuan Li, Wei Wu, and Jianguo Li. Every token counts: Generalizing 16m ultra-long context in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10208–10220, 2026c.

HuggingFace. Open llm leaderboard. 2023. URL `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

Jamba, Barak Lenz, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, et al. Jamba-1.5: Hybrid transformer-mamba models at scale. *arXiv preprint arXiv:2408.12570*, 2024.

Alexia Jolicoeur-Martineau, Rhea Sukthanker, Pashmina Cameron, and Emy Gervais. Sliding-window beats linear attention. *arXiv preprint arXiv:2608.28444*, 2026.

Greg Kamradt. Needle in a haystack - pressure testing llms. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`, 2023.

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

Kimi, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149*, 2024.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, et al. Minimax sparse attention. *arXiv preprint arXiv:2606.13392*, 2026.

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

Disen Lan, Jianbin Zheng, Yuxi Ren, Xin Xia, Xuanda Wang, Xuefeng Xiao, Xipeng Qiu, and Yu Cheng. Morphing into hybrid attention models. *arXiv preprint arXiv:2606.30562*, 2026.

Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, et al. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 3214–3252. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.229. URL https://doi.org/10.18653/v1/2022.acl-long.229.

Zhixuan Lin, Evgenii Nikishin, Xu Owen He, and Aaron Courville. Forgetting transformer: Softmax attention with a forget gate. *arXiv preprint arXiv:2503.02130*, 2025.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025a.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation. In *The Twelfth International Conference on Learning Representations*, 2024b.

Xiaoran Liu, Zhigeng Liu, Zengfeng Huang, Qipeng Guo, Ziwei He, and Xipeng Qiu. Longllada: Unlocking long context capabilities in diffusion llms. *arXiv preprint arXiv:2506.14429*, 2025b.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5):5, 2017.

Kai Lv, Xiaoran Liu, Qipeng Guo, Hang Yan, Conghui He, Xipeng Qiu, and Dahua Lin. Longwanjuan: Towards systematic measurement for long text quality. *arXiv preprint arXiv:2402.13583*, 2024.

William Merrill, Yanhong Li, Tyler Romero, Anej Svete, Caia Costello, Pradeep Dasigi, Dirk Groeneveld, David Heineman, Bailey Kuehl, Nathan Lambert, et al. Olmo hybrid: From theory to practice and back. *arXiv preprint arXiv:2604.03444*, 2026.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2381–2391. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1260. URL https://doi.org/10.18653/v1/d18-1260.

MiMo. Mimo-v2.5-pro. https://huggingface.co/collections/XiaomiMiMo/mimo-v25, 2026.

Minimax. Minimax m3: The first open-weight model with three frontier capabilities, coding & agentic frontier, 1m-context msa, native multimodality., June 2026. URL https://www.minimax.io/models/text/m3.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/V1/P16-1144. URL https://doi.org/10.18653/v1/p16-1144.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Krishna C Puvvada, Faisal Ladhak, Santiago Akle Serrano, Cheng-Ping Hsieh, Shantanu Acharya, Somshubra Majumdar, Fei Jia, Samuel Kriman, Simeng Sun, Dima Rekesh, et al. Swan-gpt: An efficient and scalable approach for long-context language modeling. *arXiv preprint arXiv:2504.08719*, 2025.

Ziqing Qiao, Yinuo Xu, Chaojun Xiao, Zhou Su, Zihan Zhou, Yingfa Chen, Xiaoyue Xu, Xu Han, and Zhiyuan Liu. Rethinking the role of efficient attention in hybrid architectures. *arXiv preprint arXiv:2606.15378*, 2026.

Zhen Qin, Weigao Sun, Dong Li, Xuyang Shen, Weixuan Sun, and Yiran Zhong. Various lengths, constant speed: Efficient language modeling with lightning attention. *arXiv preprint arXiv:2405.17381*, 2024.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *Advances in Neural Information Processing Systems*, 38:100092–100118, 2026.

Qwen. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency, August 2026a. URL https://qwen.ai/blog?id=qwen3.8-flash-next.

Qwen. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026b. URL https://qwen.ai/blog?id=qwen3.5.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URL https://doi.org/10.48550/arXiv.2311.12022.

Liliang Ren, Yang Liu, Yadong Lu, Chen Liang, Weizhu Chen, et al. Samba: Simple hybrid state space models for efficient unlimited context language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 53551–53575, 2025.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 8732–8740. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6399. URL https://doi.org/10.1609/aaai.v34i05.6399.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 4462–4472. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1454. URL https://doi.org/10.18653/v1/D19-1454.

Jianlin Su. Improving transformer: Length extrapolation ability and position robustness. 2023. URL https://spaces.ac.cn/archives/9444.

Jianlin Su. A brief history of linear attention: From imitation and innovation to feedback. Jun 2025. URL https://spaces.ac.cn/archives/11033.

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

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. Moss: An open conversational large language model. *Machine Intelligence Research*, 2024a. ISSN 2731-5398. doi: 10.1007/s11633-024-1502-8. URL https://github.com/OpenMOSS/MOSS.

Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang, Wenhui Wang, Shuming Ma, Quanlu Zhang, Jianyong Wang, and Furu Wei. You only cache once: Decoder-decoder architectures for language models. *arXiv preprint arXiv:2405.05254*, 2024b.

Zhentao Tan, Wei Chen, Jingyi Shen, Yao Liu, Xu Shen, Yue Wu, and Jieping Ye. Hydrahead: From head-level functional heterogeneity to specialized attention hybridization. *arXiv preprint arXiv:2606.20097*, 2026.

Thinking Machines. Inkling: Our open-weights model, July 2026. URL https://thinkingmachines.ai/news/introducing-inkling/.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3261–3275, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Dustin Wang, Rui-Jie Zhu, Steven Abreu, Yong Shan, Taylor Kergan, Yuqi Pan, Yuhong Chou, Zheng Li, Jibin Wu, Ge Zhang, et al. A systematic analysis of hybrid linear attention. *arXiv preprint arXiv:2507.06457*, 2025.

Pengyu Wang, Chenkun Tan, Shaojun Zhou, Qirui Zhou, Yanxin Chen, Xingyang He, Huazheng Zeng, Jijun Cheng, Chenghao Wang, Xiaomeng Qian, et al. Moss-vl technical report. *arXiv preprint arXiv:2608.15045*, 2026.

Chan Wu, Hanxiao Zhang, Lin Ju, Jinjing Huang, Youshao Xiao, Zhaoxin Huan, Siyuan Li, Fanzhuang Meng, Lei Liang, Xiaolu Zhang, et al. Rethinking memory and communication cost for efficient large language model training. *arXiv preprint arXiv:2310.06003*, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Bowen Yang, Bharat Venkitesh, Dwaraknath Gnaneshwar Talupuru, Hangyu Lin, David Cairuz, Phil Blunsom, and Acyr Locatelli. Rope to nope and back again: A new hybrid attention strategy. *Advances in Neural Information Processing Systems*, 38:64133–64157, 2026.

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

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Forty-first International Conference on Machine Learning*, 2024a.

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

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

Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, and Yoon Kim. Path attention: Position encoding via accumulating householder transformations. *arXiv preprint arXiv:2505.16381*, 2025c.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1472. URL `https://doi.org/10.18653/v1/p19-1472`.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Bo Wang, Shimin Li, Yunhua Zhou, Qipeng Guo, Xuanjing Huang, and Xipeng Qiu. Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective. *arXiv preprint arXiv:2412.14135*, 2024.

Shu Zhong, Mingyu Xu, Tenglong Ao, and Guang Shi. Understanding transformer from the perspective of associative memory. *arXiv preprint arXiv:2505.19488*, 2025.

A Extended Verification

We conduct experiments at 376M, 776M, 1B, and 3B model sizes, with the configuration detailed in Table 2. We use the Llama-3 tokenizer. All models are pretrained on the DCLM-Baseline-1.0 corpus (Li et al., 2024) in BFloat16 with a 4k context length using DeepSpeed (Rasley et al., 2020),

	376M	776M	1B	3B
Hidden size	1024	1536	2048	3072
Intermediate size	3584	5376	7168	10752
Number of layers	8	12	16	24
Attention heads	8	12	16	24
KV heads	8	12	16	24
Vocabulary size	128256	128256	128256	128256

Table 2: Model configurations used in this study.

FlashAttention2 (Dao, 2024) and FlashLinearAttention (Yang & Zhang, 2024). The 3B models use 16 NVIDIA H200 GPUs; all other models use 8 GPUs. For each size, we use a batch size of 1M tokens and pretrain for 50B tokens. We use the AdamW (Loshchilov et al., 2017) optimizer with weight decay 0.1, a maximum learning rate of 1e-3, and a cosine annealing scheduler. We use the first 0.5B tokens for warm-up, and the learning rate ends at 0.

In long-context continual pretraining, we use holistic long documents from the high-quality long-context corpus LongWanjuan (Lv et al., 2024), and continue training at a 32k context length. We follow the data mixture ratios used in Lv et al. (2024), and increase the rotary base from 10000 to 500000 for RoPE full-attention to support contexts of around 64k tokens (Liu et al., 2024b; 2025b). We use a 0.5M batch size and continue training for 5B tokens. We use the AdamW (Loshchilov et al., 2017) optimizer with 0 weight decay, a peak learning rate of 1e-3, and a cosine annealing scheduler. We use the first 1B tokens for warm-up, and finally decay the learning rate to 0.

A.1 Supplementary Results

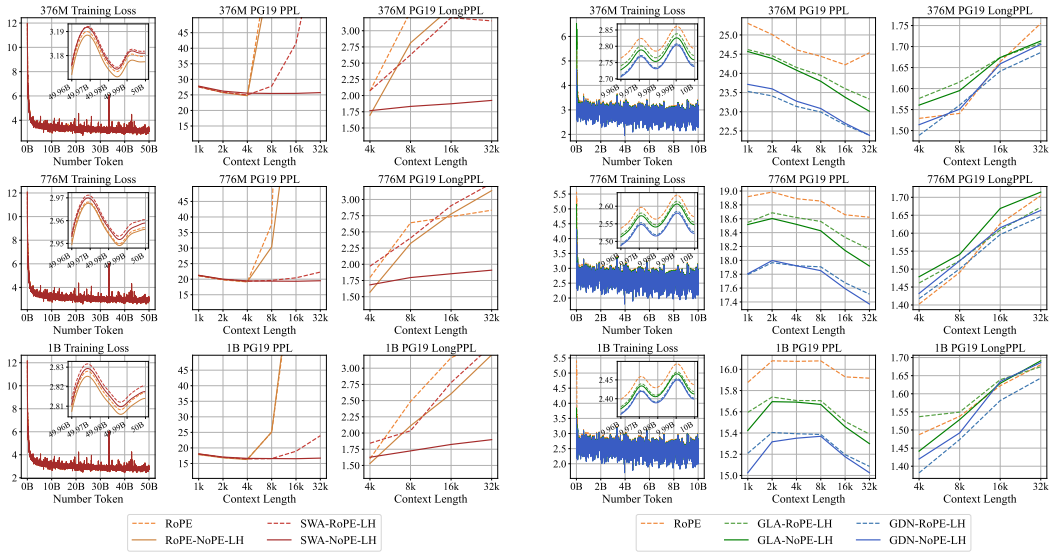
Here, we include additional figures, referenced but not displayed in the main text. Figure 11 and 12 illustrate the training dynamics, perplexity, and LongPPL of head-wise and layer-wise hybrid models, showing that the hybrid models achieve lower training loss, with the NoPE hybrid models demonstrating superior extrapolation capabilities. Figures 13 and 12 show the long-context performance of head-wise models, while Figure 15 compares the length extrapolation and context extension capabilities of head-wise hybrid models; notably, the Seesaw Effect is less pronounced in head-wise hybrids than in layer-wise hybrids. Figure 16 highlights the extrapolation advantages of the RoPE-NoPE hybrid model, demonstrating that Yang et al. (2026) yields better results than the NTK method, which inspires the length extrapolation of LA hybrid models and the proposal of the Matthew effect of length extrapolation in hybrid position. Figures 17 and 18 illustrate the extrapolation performance of GLA-NoPE and GDN-NoPE models using SWLA and log-scaled NoPE. Finally, Figure 19 compares the extrapolation performance of RoPE-NoPE, SWA-NoPE, GLA-NoPE, and GDN-NoPE using these methods, showing that these models achieve comparable results.

A.2 More Verification of Seesaw Effect

To verify that the seesaw effect is stable, we first compare long-context continual pretraining results of SWA and LA hybrids under different peak learning rates, including 1e-4, 2e-4, 5e-4, and 1e-3, where 1e-3 is the default learning rate. We conduct this comparison at both 376M and 776M scales. The results are shown in Figure 20, and the point with a zero learning rate denotes the direct length extrapolation result before long-context continual pretraining. The seesaw effect remains stable under different learning rates: after context extension, SWA-NoPE hybrids, especially on long-context tasks, consistently lag behind LA-NoPE hybrids. We also verify the existence of the Seesaw Effect in a YOCO-like layer-wise hybrid models, in which all efficient-attention layers are placed at the lower layers. The results are shown in Figure 21. This further confirms that the Seesaw Effect is a general phenomenon in SWA and LA hybrids.

A.3 Broader Discussion

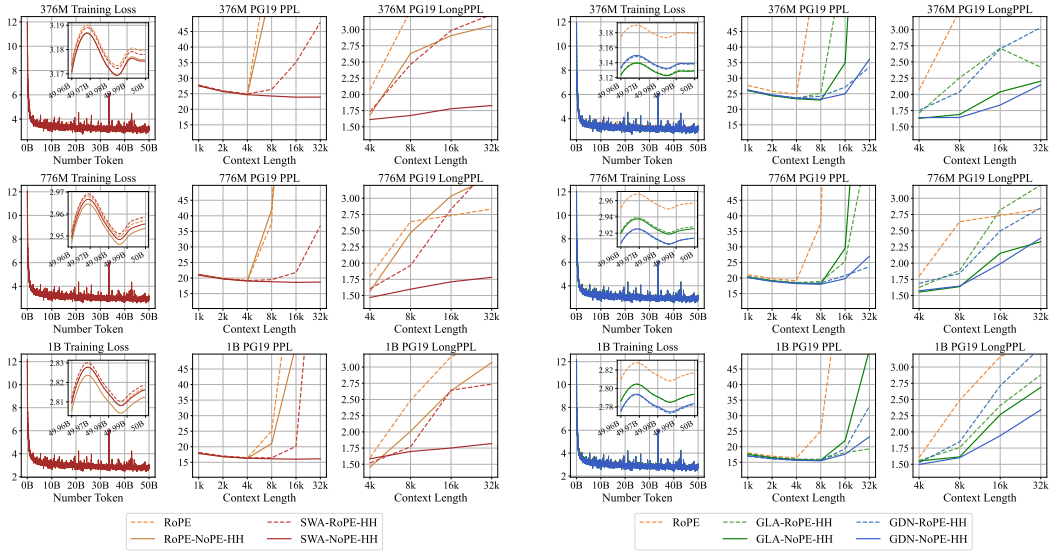
Smaller Rotary Base in SWA Hybrids We have compared short-context training results for SWA hybrids with different window sizes in short-context pretraining and long-context pretraining above. During window-size extension, the RoPE base must be adjusted accordingly. When the window size



(a) Results on Full Attention and SWA Hybrid.

(b) Results on Full Attention and GLA/GDN Hybrid.

Figure 11: Training loss, PPL, and LongPPL among different layer-wise hybrids after 4k context length pretraining. SWA/GLA/GDN-NoPE hybrids achieve stronger long-context language modeling.



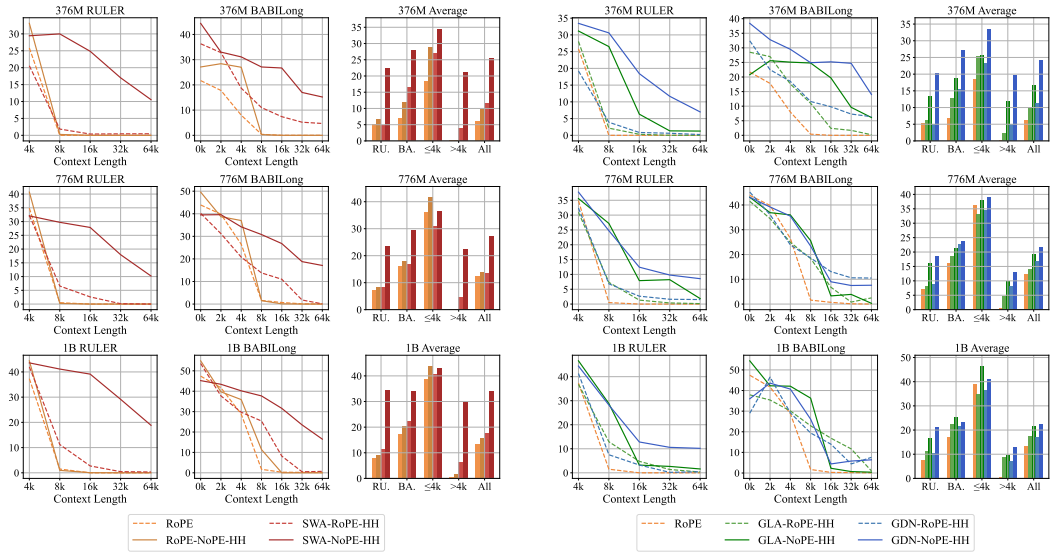
(a) Results on Full Attention and SWA Hybrid.

(b) Results on Full Attention and GLA/GDN Hybrid.

Figure 12: Training loss, PPL and LongPPL among different head-wise hybrids after 4k context length pretraining. SWA/GLA/GDN-NoPE hybrids achieve stronger long-context language modeling.

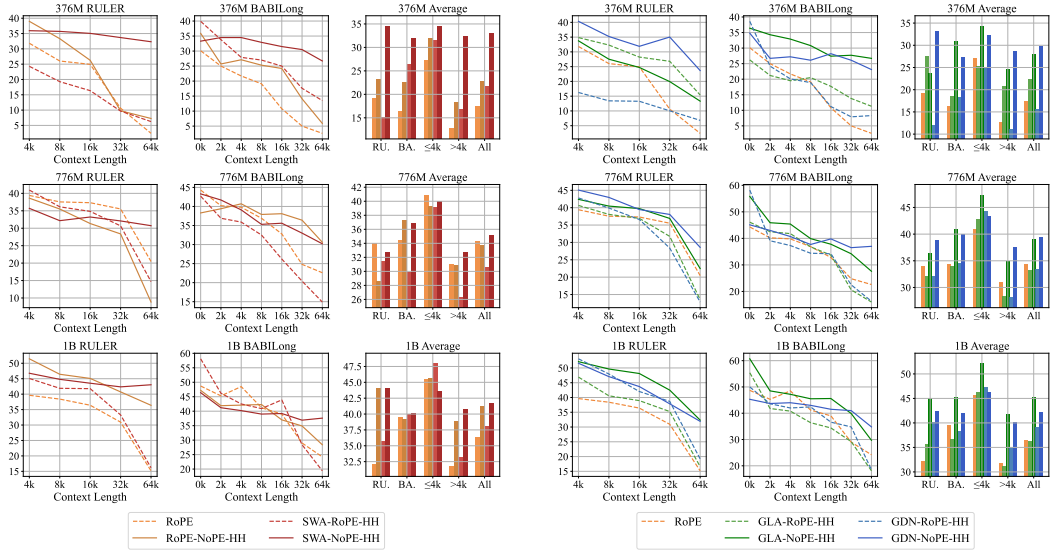
is increased from 128 to 256, 512, 1024, 2048, and 4096, the corresponding RoPE bases, calculated according to the scaling laws for RoPE extrapolation (Liu et al., 2024b) and rounded up, become 90000, 700000, 6000000, 50000000, and 400000000, respectively.

These values are clearly very large. This is because the initial window size is only 128, meaning that only a very limited number of dimensions have observed the complete positional information range; consequently, substantial interpolation is required. However, an excessively large RoPE base can reduce the discriminability of the positional embedding. This motivated us to investigate the use of smaller RoPE bases in SWA. We kept the window size fixed at 128 and experimented with



(a) Results on Full Attention and SWA Hybrid. (b) Results on Full Attention and GLA/GDN Hybrid.

Figure 13: Comparison of long-context performance under RULER and BABILong benchmarks among different head-wise hybrid models after 4k context length pretraining.

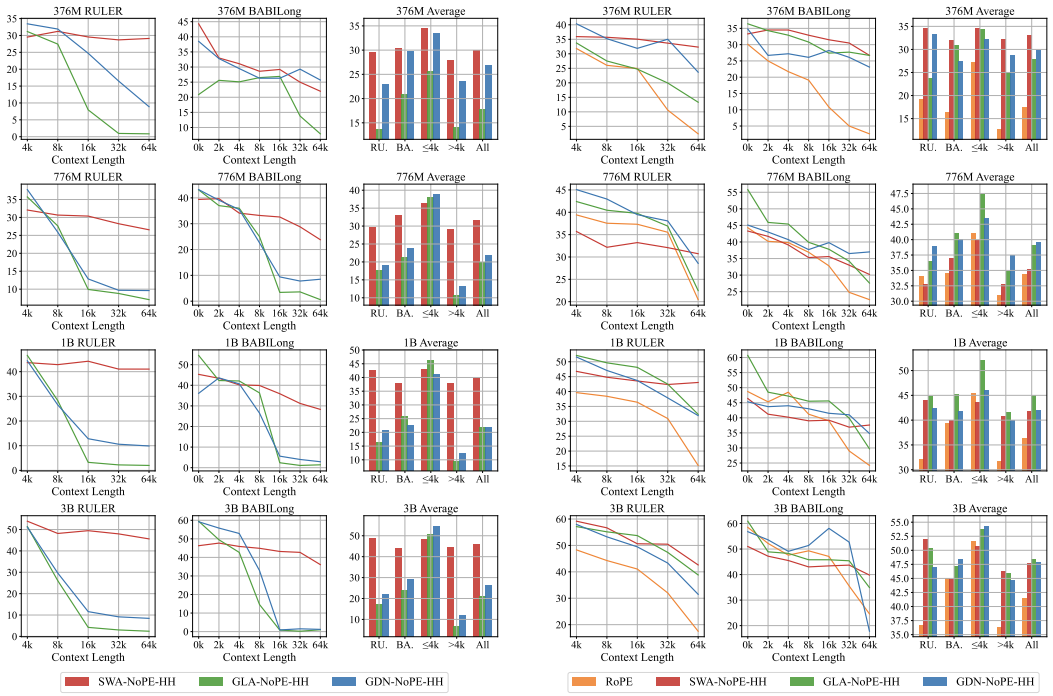


(a) Results on Full Attention and SWA Hybrid. (b) Results on Full Attention and GLA/GDN Hybrid.

Figure 14: Comparison of long-context performance under RULER and BABILong benchmarks among different head-wise hybrid models after 32k context length continual pretraining.

RoPE bases ranging from 5,000 to 100 on both the 376M and 776M models. The results confirm that smaller bases can lead to better length extrapolation and context extension performance as shown in Table 18, but we do not observe a consistent trend across all settings. So we do not emphasize this finding in the main text. We will further investigate the underlying mechanism and try to develop an optimal scheduler for the training length, window size, and RoPE base in SWA hybrids.

Performance Difference within LA Hybrids In the main text, we analyze LA hybrids by treating GLA and GDN as a whole. Nevertheless, there still exist differences within LA hybrids. For example, in short-context training, GLA hybrids are better at fitting but weaker at extrapolation than GDN



(a) Performance of log-scale NoPE extrapolation.

(b) Performance after long-context pretraining.

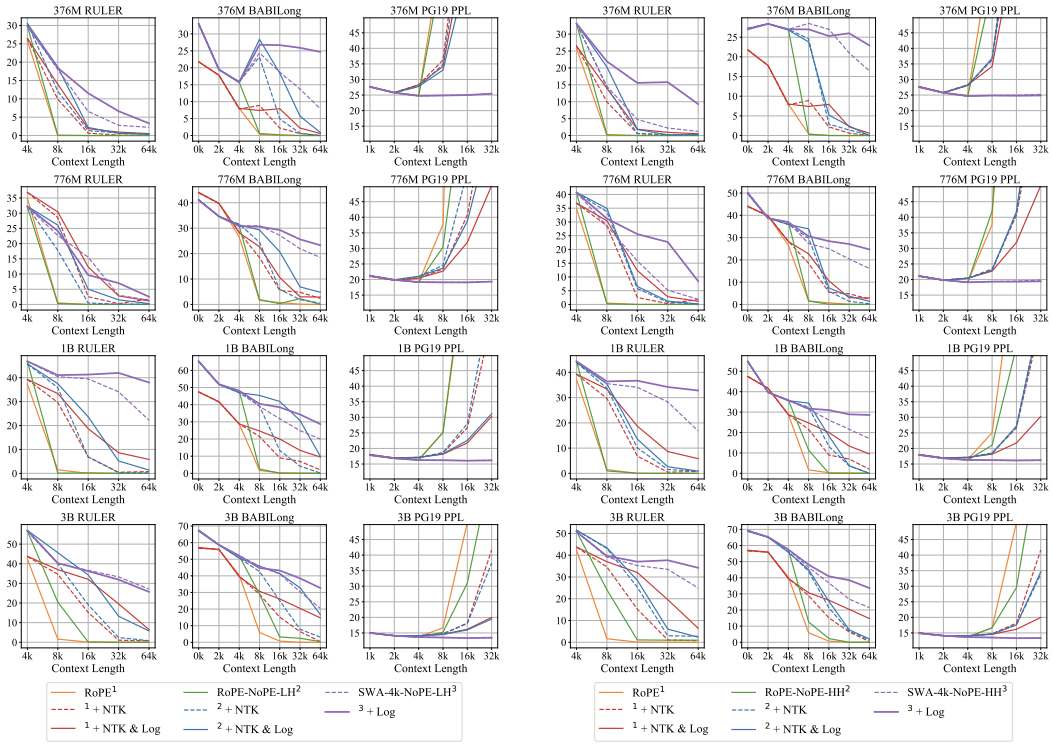
Figure 15: Comparison of the long-context performance of the SWA/GLA/GDN-NoPE head-wise hybrid models under log-scale NoPE extrapolation after short-context pretraining (left) and after long-context continual pretraining (right).

hybrids, as shown in Figures 1, 2, 13 and 14. In contrast, GDN's fitting and extrapolation performance are more moderate, falling between GLA and SWA. However, when combined with SWLA, GLA hybrids show better length extrapolation performance than GDN hybrids, as shown in Figures 17 and 18. There might exist a mechanism related to the introduction of the delta rule (Yang et al., 2024b), and we intend to conduct additional analysis in the future.

Transfer Training Experiment Inspired by the effectiveness of hybrid position, we also attempt to transfer SWA/GLA/GDN-RoPE hybrids trained in 4k short contexts to corresponding SWA/GLA/GDN-NoPE in the long-context training phase, namely dropping RoPE in long-context training like DRoPE (Gelberg et al., 2026; Merrill et al., 2026). We report both short- and long-context evaluation results in Table 5, 6, 11, 12 and 17, and find no consistent pattern across models or sizes. This area may still need further exploration.

A.4 Efficiency Comparison

Additionally, we include a brief comparison of inference and training efficiency, as shown in Figure 22 and 23. Regarding inference efficiency, we compare traditional full-attention-only models with 3:1 layer-wise or head-wise hybrid models that combine SWA, GLA, or GDN with NoPE full attention, across model sizes of 376M, 776M, 1B, and 3B. We measure prefilling efficiency with Time to First Token (TTFT) and decoding efficiency with Time per Output Token (TPOT), both in microseconds. We evaluate memory efficiency based on peak allocated GPU memory and cache size, both measured in GB. As shown in Figure 22, hybrid models offer clear advantages as the model size and context length increase. Notably, although head-wise hybrids compute attention twice in every layer, the advantage in larger model sizes and longer context lengths is still remarkable. Similarly, regarding training efficiency, we compare training throughput on 8 H200 GPUs, measured by TGS (tokens per GPU per second) in 4k and 32k context lengths. We still compare full-attention-only models with 3:1 layer-wise or head-wise hybrid models on 376M, 776M, and 1B, without 3B, since training 3B



(a) Results on Layer-wise Hybrid Model (LH). (b) Results on Head-wise Hybrid Model (HH).

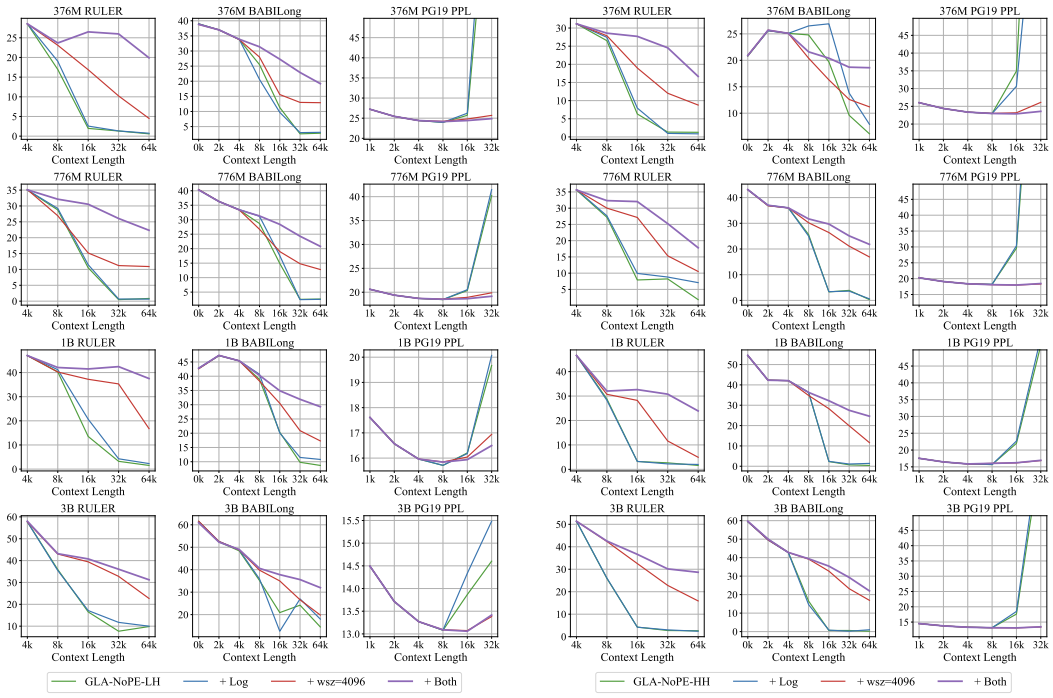
Figure 16: Comparison of length extrapolation between RoPE-only and RoPE-NoPE hybrid full-attention models. Both layer-wise and head-wise RoPE-NoPE hybrids have better long-context performance and can be extrapolated effectively by sliding-window RoPE attention and log-scaled NoPE attention, better than traditional interpolation method such as NTK.

models requires 16 H200 GPUs. As shown in Figure 23, hybrid models still offer a larger throughput, especially for the 1B model size and 32k context length.

Moreover, because the GLA and GDN store their recurrent states in higher FP32 precision, their cache sizes are comparable to the cache overhead of the SWA hybrid with a window size of 128 and FP16 precision, so their curves overlap in the memory-efficiency comparison. In addition, because the QKV matrices of head-wise SWA hybrids are concatenated rather than split, as in head-wise LA hybrids, the intermediate memory footprint is larger, so head-wise SWA hybrids share the same peak inference memory as full attention and offer lower training throughput than LA hybrids.

A.5 More Related Work

Beyond efficiency, previous work also suggests that hybrid models can bring performance gains within the training context length (Kimi et al., 2026; Merrill et al., 2026; Yang et al., 2026). Early works (Fu et al., 2022; Cooper et al., 2026) show that hybrid models can combine the advantages of traditional softmax attention in retrieval with the advantages of linear attention in state tracing. Kimi-K3 (Kimi et al., 2026) and OLMo Hybrid (Merrill et al., 2026) further verify at larger scales that LA hybrids can improve model expressivity and pretraining scaling efficiency. In addition to layer-wise hybrids, Hymba (Dong et al., 2024) and HydraHead (Tan et al., 2026) also verify the effectiveness of head-wise hybrids. Beyond empirical validation, some works have begun to analyze SWA and LA hybrids theoretically. For example, Wang et al. (2025) suggests that a ratio around 3:1 suits hybrid models, while Wu et al. (2023) points out that retrieval is mainly handled by full attention, whereas SWA or LA affects the training trajectory.



(a) Results on GLA-NoPE-LH.

(b) Results on GLA-NoPE-HH.

Figure 17: Comparison of length extrapolation in GLA-NoPE hybrids. Applying a sliding window for GLA and a log scale for NoPE attention achieves the best length extrapolation performance.

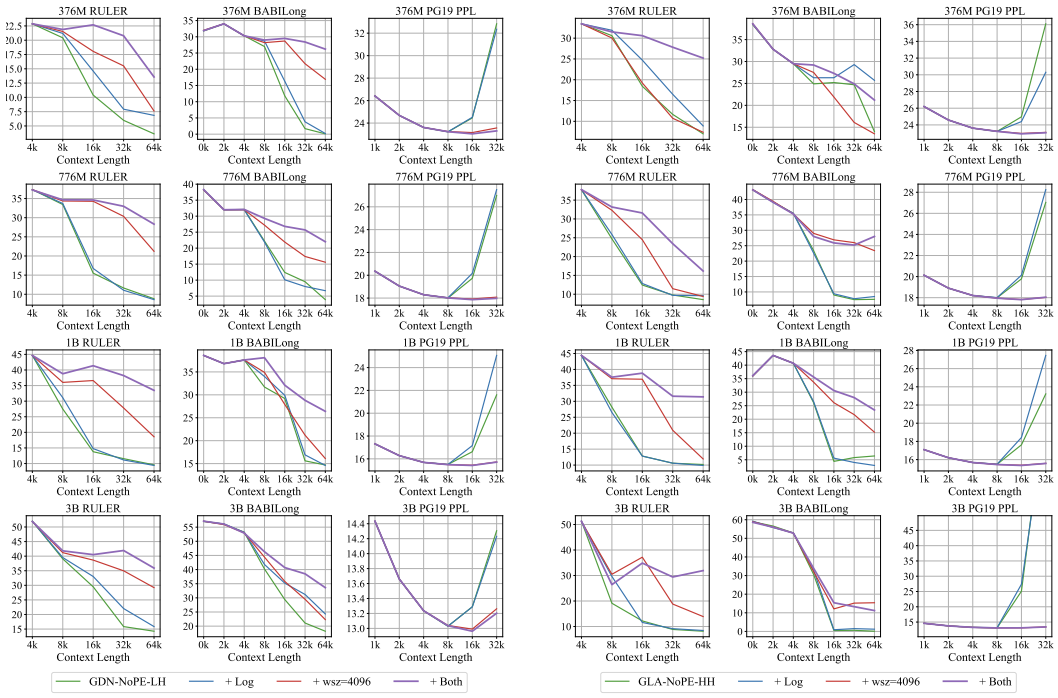
Besides improving long-context performance within the training length, hybrid models have also been shown to improve length extrapolation, or length generalization (Press et al., 2022). For example, iRoPE (AI, 2025), RNoPE (Yang et al., 2026), and SWAN-GPT (Puvvada et al., 2025) use RoPE only in some layers and restrict the attention window of these layers during extrapolation, while keeping global position awareness only in NoPE layers, achieving strong extrapolation performance. This hybrid position design has also proven to improve retrieval performance within the training length (Yang et al., 2026). Similarly, in LA hybrids, Jamba (Jamba et al., 2024), Kimi-K3 (Kimi et al., 2026), and HypeNet (Chen et al., 2026) find that RoPE can also be removed from full-attention layers, which also helps direct length generalization (Press et al., 2022).

Although LA hybrids do not use explicit positional embeddings, this does not mean there is no positional bias. Linear attention also contains various position designs, some of which have already been used to improve traditional softmax attention and build stronger position embeddings or attentions (Su, 2025), such as FoX (Lin et al., 2025), DeltaFormer (Zhong et al., 2025), and PaTH (Yang et al., 2025c). Accordingly, we can extrapolate LA hybrids in the same way as RoPE-NoPE hybrids by limiting position-biased attention to local windows while strengthening global perception in modules without positional bias, which is also one of the contributions in this paper.

A.6 Limitations

This paper provides a mechanistic analysis of long-context hybrid models, with a broad training-validation-observation-induction loop. Even so, several directions remain open.

- **Validation scope**. Our current validation pipeline is focused on context extension and length extrapolation in the pretraining stage. We still need to extend it to post-training (Zeng et al., 2024) and multimodal, especially vision-language (Yang et al., 2025a; Wang et al., 2026) models. On that basis, we should also extend our evaluation to complex reasoning, agentic tasks, multimodal retrieval, and multimodal understanding.



(a) Results on Layer-wise Hybrid Model (LH).

(b) Results on Head-wise Hybrid Model (HH).

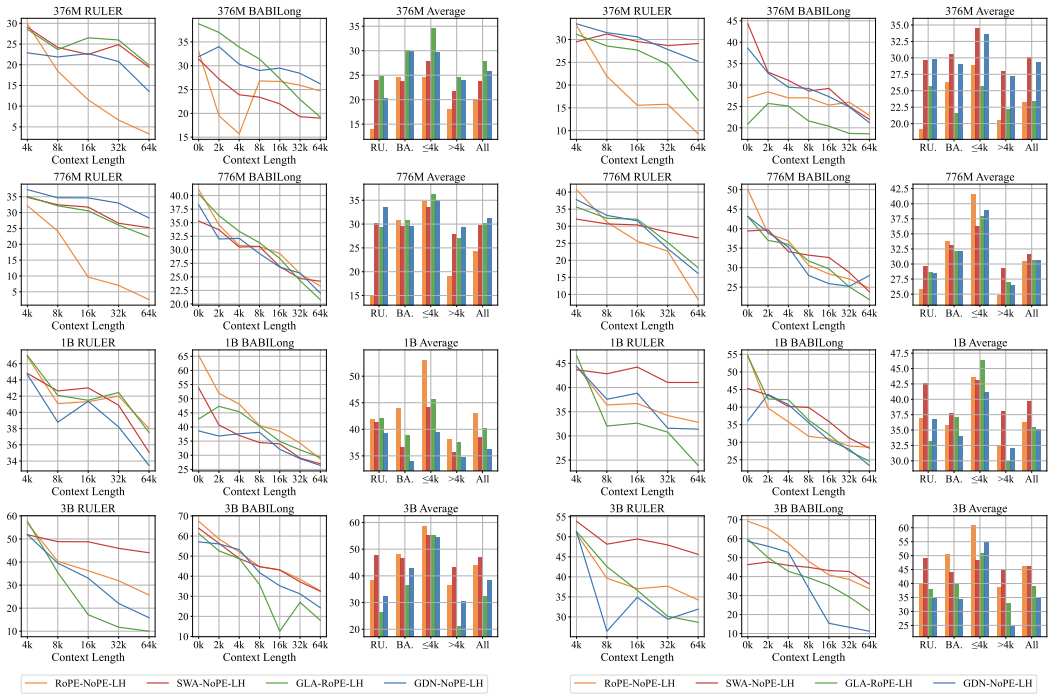
Figure 18: Comparison of length extrapolation in GDN-NoPE hybrids. Applying a sliding window for GDN and a log scale for NoPE attention achieves the best length extrapolation in most cases.

- **Architecture scope**. Our analysis focuses on RoPE/NoPE hybrids, SWA hybrids, and LA hybrids. We should apply the same mechanistic lens to other hybrid families, such as sparse-attention (Liu et al., 2024a; Lai et al., 2026; Qwen, 2026a) and compressed-attention (Xu et al., 2026; Chu et al., 2026). We leave this broader comparison for future work.
- **Mechanism scope**. Even within the hybrid families studied here, we only probe a subset of position and attention design choices. We will also further our study by analyzing techniques such as sink bias (Xiao et al., 2024; Agarwal et al., 2025; MiMo, 2026), partial RoPE (Zeng et al., 2026; Qwen, 2026b), gated attention variants (Qiu et al., 2026; Qwen, 2026b), and ablation of short convolution (Yang et al., 2025b;c; Thinking Machines, 2026).

B Detailed Evaluation Results

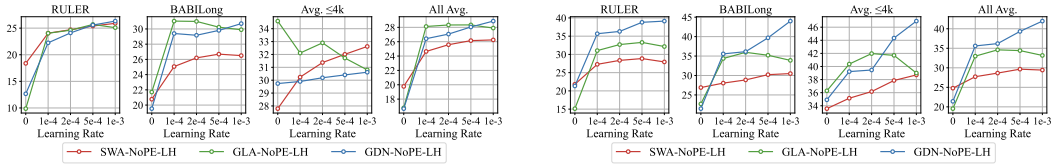
In this section, we first report the results of short-context evaluation. We evaluate both short-context and long-context models on downstream tasks mainly in Open LLM Leaderboard (HuggingFace, 2023), including TruthfulQA (Lin et al., 2022), LAMBADA(Paperno et al., 2016), PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2020), ARC (Clark et al., 2018), GPQA (Rein et al., 2023), SocialIQA (Sap et al., 2019), OpenBookQA (Mihaylov et al., 2018), and SuperGLUE (Wang et al., 2019). All models are tested within a 2k context length. The results are reported in Tables 3, 4, 5 and 4. Hybrid models achieve performance comparable to the traditional RoPE-based full-attention-only model, before and after long-context continual pretraining.

Regarding the long-context evaluation analyzed before, we report the long-context performance of hybrid models after short-context pretraining in Tables 7, 10, 9 and 10. The length extrapolation performance validates the Matthew Effect of Length Extrapolation and verifies the effectiveness of sliding-window linear attention. We also report long-context performance of hybrid models after long-context continual pretraining in Tables 11 and 12. The context extension performance validates the Seesaw Effect. We further detail the long-context performance of SWA layer-wise



(a) Results on Layer-wise Hybrid Model (LH). (b) Results on Head-wise Hybrid Model (HH).

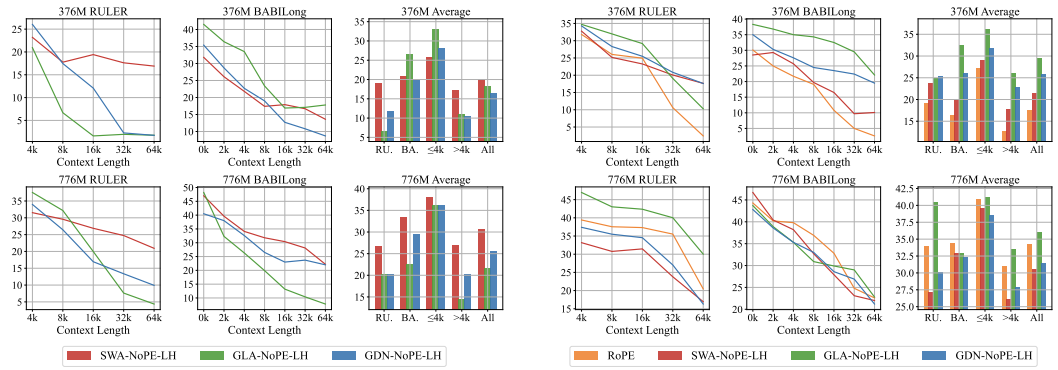
Figure 19: Comparison of length extrapolation with Matthew effect in RoPE-NoPE, GLA-NoPE, and GDN-NoPE hybrids as well as SWA-NoPE hybrids.



(a) Results of 376M models. (b) Results of 776M models.

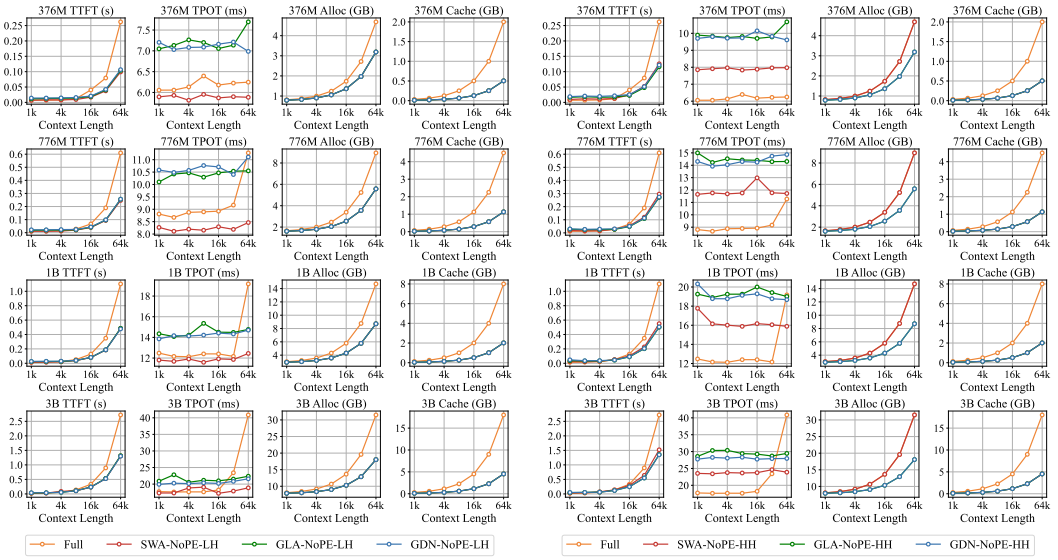
Figure 20: Comparison of long-context training performance between SWA-NoPE-LH, GLA-NoPE-LH, and GDN-NoPE-LH under different learning rates, where 0 learning rate means the performance before long-context continual pretraining.

hybrid models with different window sizes in short-context pretraining and long-context continual pretraining in Tables 13 and 14, validating Short-Window Weariness and Long-Window Laziness. Besides, in Table 15, we also report the long-context performance of SWA layer-wise hybrid models after long-context continual pretraining with extended window sizes (wsz) and LongCE enhancement, both relieving the short-context learning trap effectively. Additionally, in Table 17, we verify the existence of Seesaw Effect under different layer-wise hybrid layout. Finally, in Table 18, we report the long-context performance of SWA-NoPE-LH with different rotary base in SWA.



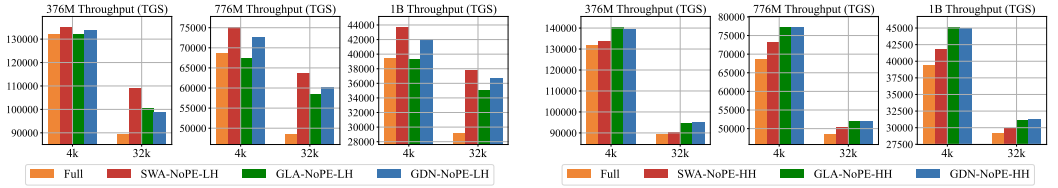
(a) Performance of log-scale NoPE extrapolation. (b) Performance after long-context pretraining.

Figure 21: Comparison of the long-context performance of the SWA/GLA/GDN-NoPE layer-wise hybrid models in YOCO-like layer layout under log-scale NoPE extrapolation after short-context pretraining (left) and after long-context pretraining (right). The Seesaw Effect still exists.



(a) Results on Layer-wise Hybrid Model (LH). (b) Results on Head-wise Hybrid Model (HH).

Figure 22: Comparison of full attention model and hybrid models on inference efficiency.



(a) Results on Layer-wise Hybrid Model (LH). (b) Results on Head-wise Hybrid Model (HH).

Figure 23: Comparison of full attention model and hybrid models on training efficiency.

	TQA	LMD	PIQA	Hella	Wino	ARC-e	ARC-c	GPQA	SIQA	OBQA	SG	MMLU	Avg.
376M Short													
RoPE	36.8	40.2	65.9	33.8	52.2	**39.7**	26.4	26.3	39.2	27.4	42.1	24.6	37.9
RoPE-NoPE-LH	36.4	41.7	65.9	33.8	51.1	39.2	**27.5**	**28.8**	40.3	25.8	**45.2**	**26.9**	**38.5**
SWA-RoPE-LH	35.8	42.6	65.8	33.4	52.3	37.0	25.4	24.8	39.9	**27.6**	44.8	25.9	37.9
SWA-NoPE-LH	36.4	42.0	66.4	33.4	50.2	37.2	24.1	25.8	39.7	21.6	43.6	25.6	37.2
GLA-RoPE-LH	**37.0**	42.7	66.4	33.8	52.3	38.8	25.1	24.2	39.0	26.6	44.7	26.3	38.1
GLA-NoPE-LH	36.0	44.2	65.9	34.1	52.4	**39.7**	23.4	23.7	38.9	27.4	44.9	24.9	38.0
GDN-RoPE-LH	36.1	**45.3**	**66.5**	**34.4**	**52.9**	38.8	24.1	26.8	**40.6**	**27.6**	43.6	25.0	**38.5**
GDN-NoPE-LH	36.0	44.9	65.1	**34.4**	50.5	37.7	25.4	**28.8**	39.8	27.2	43.5	26.4	38.3
776M Short													
RoPE	35.8	51.3	68.4	42.0	52.7	**45.9**	27.1	25.3	40.6	26.6	43.0	24.6	40.3
RoPE-NoPE-LH	36.4	50.9	68.8	41.5	52.0	41.3	26.4	27.3	40.4	24.6	43.2	25.1	39.8
SWA-RoPE-LH	36.7	49.8	**69.9**	42.2	51.6	44.4	30.5	**28.3**	40.7	27.4	45.2	25.2	41.0
SWA-NoPE-LH	36.3	50.7	69.8	41.7	53.9	42.3	28.8	24.2	40.9	26.6	**45.7**	26.4	40.6
GLA-RoPE-LH	36.8	52.1	69.2	42.2	53.7	45.2	29.5	26.8	**41.0**	**27.8**	43.5	25.6	41.1
GLA-NoPE-LH	37.3	52.6	69.8	42.5	52.6	41.3	31.2	23.7	40.1	24.2	43.8	24.9	40.3
GDN-RoPE-LH	**37.7**	**53.9**	68.9	42.2	53.8	42.2	**32.2**	19.2	40.4	25.6	44.6	25.9	40.5
GDN-NoPE-LH	37.3	52.6	69.4	**42.9**	**54.6**	42.0	29.8	27.8	40.8	26.4	45.4	**26.5**	**41.3**
1B Short													
RoPE	**38.7**	56.5	70.8	46.8	55.3	47.6	29.5	**29.3**	41.3	22.0	44.5	**26.8**	42.4
RoPE-NoPE-LH	36.8	57.4	71.6	**48.6**	55.3	**48.9**	28.8	21.2	40.7	23.6	43.2	26.1	41.8
SWA-RoPE-LH	37.4	57.9	70.2	47.6	53.4	46.4	31.5	24.2	41.7	23.4	44.3	26.2	42.0
SWA-NoPE-LH	37.9	57.1	71.3	46.8	52.8	47.3	**31.9**	23.7	41.1	**25.8**	45.3	26.6	42.3
GLA-RoPE-LH	37.4	57.4	71.0	**48.6**	56.3	47.1	31.5	25.3	41.6	23.2	43.1	26.4	42.4
GLA-NoPE-LH	36.7	59.1	**71.9**	48.4	55.2	46.4	**31.9**	23.2	41.0	25.6	44.5	25.1	42.4
GDN-RoPE-LH	37.9	**59.2**	71.8	47.9	**56.6**	47.6	29.8	21.7	**41.8**	22.2	46.6	25.2	42.4
GDN-NoPE-LH	37.7	57.7	71.2	47.2	55.0	46.4	30.9	28.3	41.5	24.6	**47.4**	23.4	**42.6**
3B Short													
RoPE	36.6	62.9	73.3	55.2	58.0	51.7	**33.6**	25.8	41.7	**27.6**	45.4	25.6	44.8
RoPE-NoPE-LH	36.0	63.5	73.6	**56.1**	58.0	52.6	29.8	21.2	**42.6**	**27.6**	46.5	**26.7**	44.5
SWA-RoPE-LH	36.6	62.9	73.6	55.3	58.6	54.7	31.9	25.3	42.3	22.0	47.8	26.0	44.7
SWA-NoPE-LH	37.3	62.2	73.5	55.7	57.3	53.6	32.5	21.7	42.0	23.6	45.0	26.5	44.3
GLA-RoPE-LH	**37.6**	62.3	73.1	52.8	56.9	49.9	32.5	25.8	42.0	25.4	46.7	25.6	44.2
GLA-NoPE-LH	36.8	64.2	73.7	54.5	57.4	50.3	33.2	**29.3**	42.0	26.2	47.7	25.0	45.0
GDN-RoPE-LH	37.0	**65.4**	73.7	54.8	**60.6**	53.8	32.2	22.7	41.6	26.2	**48.6**	26.4	**45.3**
GDN-NoPE-LH	36.1	64.2	**74.2**	**56.1**	57.7	**55.7**	30.9	21.2	42.4	27.0	46.8	26.5	44.9

Table 3: Short-context performance of layer-wise hybrid models after short-context pretraining.

	TQA	LMD	PIQA	Hella	Wino	ARC-e	ARC-c	GPQA	SIQA	OBQA	SG	MMLU	Avg.
376M Short													
RoPE	36.8	40.2	65.9	33.8	52.2	39.7	**26.4**	26.3	39.2	27.4	42.1	24.6	37.9
RoPE-NoPE-HH	**37.1**	40.1	66.0	33.3	50.4	40.0	24.8	27.8	39.9	27.8	43.3	25.8	38.0
SWA-RoPE-HH	35.1	41.3	65.4	33.6	51.3	39.7	23.7	22.7	38.8	**28.0**	45.0	26.3	37.6
SWA-NoPE-HH	35.8	41.7	65.5	33.8	51.7	39.0	**26.4**	27.8	38.7	23.4	42.2	25.2	37.6
GLA-RoPE-HH	36.0	**46.0**	66.0	**35.0**	52.4	**40.2**	24.8	25.8	39.1	23.6	43.5	25.6	38.2
GLA-NoPE-HH	36.0	45.2	65.7	34.6	51.5	37.7	25.8	27.8	39.6	27.8	44.7	26.2	38.5
GDN-RoPE-HH	36.8	44.9	66.3	34.7	51.3	38.3	24.8	26.8	**40.6**	**28.0**	44.0	**26.7**	38.6
GDN-NoPE-HH	35.8	44.2	**67.1**	34.7	**52.6**	39.7	23.4	**30.8**	39.1	26.4	**46.1**	25.4	**38.8**
776M Short													
RoPE	35.8	51.3	68.4	42.0	52.7	45.9	27.1	25.3	40.6	26.6	43.0	24.6	40.3
RoPE-NoPE-HH	35.2	51.6	68.9	42.4	54.3	42.2	27.1	24.8	40.1	21.2	43.6	25.3	39.7
SWA-RoPE-HH	37.3	49.7	67.9	41.5	52.0	40.7	26.4	24.2	39.6	18.8	42.8	26.1	38.9
SWA-NoPE-HH	35.8	50.8	69.7	42.3	53.4	42.2	26.8	25.3	40.0	24.8	**44.3**	**26.4**	40.1
GLA-RoPE-HH	**38.3**	51.9	70.2	42.4	52.3	43.7	29.8	**26.3**	**40.9**	**31.6**	42.9	25.9	**41.4**
GLA-NoPE-HH	36.0	53.0	70.1	42.9	52.3	45.5	28.1	24.8	39.7	23.2	44.0	25.3	40.4
GDN-RoPE-HH	36.4	52.0	**70.7**	**43.5**	53.0	**46.4**	**30.2**	24.2	40.7	21.8	44.2	25.5	40.7
GDN-NoPE-HH	37.6	**54.0**	68.8	43.2	**55.2**	45.7	29.5	21.2	40.3	27.0	43.3	24.2	40.8
1B Short													
RoPE	**38.7**	56.5	70.8	46.8	55.3	47.6	29.5	**29.3**	41.3	22.0	44.5	**26.8**	42.4
RoPE-NoPE-HH	36.6	57.0	70.0	47.6	55.0	46.4	30.2	25.8	41.2	**27.8**	45.9	25.1	42.4
SWA-RoPE-HH	**38.7**	56.3	70.6	47.2	55.3	46.6	**32.2**	27.3	40.9	24.6	44.1	25.9	42.5
SWA-NoPE-HH	36.1	56.1	71.1	47.9	55.0	46.6	30.9	22.2	41.0	26.8	45.1	26.3	42.1
GLA-RoPE-HH	36.1	58.8	70.6	**49.6**	54.5	**48.9**	31.9	20.2	41.7	26.6	43.3	26.3	42.4
GLA-NoPE-HH	36.7	**59.1**	71.0	48.7	**55.8**	47.8	**32.2**	24.8	41.3	26.6	44.6	24.9	**42.8**
GDN-RoPE-HH	36.1	57.9	**72.0**	**49.6**	54.7	46.0	29.8	21.7	**41.8**	27.2	46.7	26.6	42.5
GDN-NoPE-HH	35.2	58.8	70.1	49.3	53.6	47.4	31.5	20.7	**41.8**	19.6	**47.8**	25.7	41.8
3B Short													
RoPE	36.6	62.9	73.3	55.2	58.0	**51.7**	33.6	25.8	41.7	27.6	45.4	25.6	44.8
RoPE-NoPE-HH	36.3	63.5	73.1	56.0	57.5	49.9	32.5	28.3	42.0	23.8	46.9	24.4	44.5
SWA-RoPE-HH	37.1	63.2	73.7	56.3	58.4	51.2	35.3	26.8	42.4	25.8	46.4	**27.6**	45.3
SWA-NoPE-HH	36.4	62.4	73.2	54.4	58.5	51.5	31.5	21.7	42.3	22.4	45.9	25.7	43.8
GLA-RoPE-HH	36.6	63.3	73.7	56.2	59.0	51.0	31.9	24.2	43.1	20.8	46.4	27.4	44.5
GLA-NoPE-HH	35.4	63.9	74.2	56.3	58.3	50.6	31.2	24.2	42.7	**28.8**	**48.6**	25.9	45.0
GDN-RoPE-HH	**37.7**	**64.1**	**74.5**	**57.5**	59.8	51.2	34.9	27.3	**43.5**	24.0	45.0	25.6	45.4
GDN-NoPE-HH	36.8	63.9	73.7	54.7	**60.3**	50.8	**37.3**	**29.3**	42.6	25.8	48.1	24.8	**45.7**

Table 4: Short-context performance of head-wise hybrid models after short-context pretraining.

	TQA	LMD	PIQA	Hella	Wino	ARC-e	ARC-c	GPQA	SIQA	OBQA	SG	MMLU	Avg.
376M Long													
RoPE	35.2	35.8	64.3	31.2	51.2	36.7	25.8	24.8	38.5	28.0	43.6	24.2	36.6
RoPE-NoPE-LH	36.0	38.7	64.2	31.7	52.5	37.6	**28.8**	25.8	**39.2**	**28.2**	42.9	24.7	37.5
SWA-RoPE-LH	35.4	39.3	64.2	31.0	52.4	39.2	25.8	23.7	38.3	28.0	**45.3**	25.2	37.3
+ DRoPE	35.5	39.1	**65.0**	30.8	51.6	39.0	23.4	24.8	38.1	27.6	44.7	25.7	37.1
SWA-NoPE-LH	35.8	39.6	64.3	31.4	49.0	36.7	25.4	25.3	37.6	26.4	44.4	25.0	36.7
GLA-RoPE-LH	37.1	41.6	64.4	32.0	52.2	37.0	24.1	25.3	39.1	27.2	44.9	**26.2**	37.6
+ DRoPE	36.0	40.2	63.7	31.6	51.2	37.6	25.1	**26.3**	38.4	27.4	43.4	25.7	37.2
GLA-NoPE-LH	**37.6**	41.6	64.4	31.7	51.5	38.6	23.4	25.8	38.0	27.8	44.3	24.6	37.4
GDN-RoPE-LH	36.0	**43.7**	63.8	**32.2**	**52.8**	39.0	23.4	25.8	38.7	27.6	44.2	24.8	**37.7**
+ DRoPE	35.8	43.3	64.1	32.1	52.0	**39.3**	22.7	24.2	39.0	27.6	44.7	25.1	37.5
GDN-NoPE-LH	36.4	41.9	64.5	32.1	51.2	37.6	25.4	24.2	39.0	28.0	44.0	24.9	37.4
776M Long													
RoPE	**37.7**	47.9	66.1	38.7	52.6	41.1	26.1	**25.8**	39.6	27.6	43.1	24.6	39.2
RoPE-NoPE-LH	37.0	48.3	67.0	38.4	51.8	41.5	26.8	**25.8**	40.3	27.6	42.9	24.6	39.3
SWA-RoPE-LH	36.4	45.1	67.7	38.6	51.9	**42.2**	25.4	17.7	39.0	27.6	45.3	24.0	38.4
+ DRoPE	36.3	45.7	67.8	38.7	52.0	41.8	25.8	21.2	39.0	26.4	44.9	23.9	38.6
SWA-NoPE-LH	37.0	45.8	66.7	38.3	52.3	41.5	28.1	24.8	39.8	**27.8**	**45.5**	24.3	39.3
GLA-RoPE-LH	37.6	49.1	67.7	39.0	53.8	41.6	27.8	24.2	40.3	26.8	**45.5**	24.7	39.8
+ DRoPE	37.3	49.7	**68.0**	38.7	**54.6**	42.0	25.8	24.2	40.1	27.2	45.0	25.2	39.8
GLA-NoPE-LH	36.8	50.6	67.2	39.2	52.3	41.8	**30.2**	20.7	40.0	27.6	45.0	25.6	39.7
GDN-RoPE-LH	37.4	**51.4**	66.9	**40.1**	53.5	42.0	29.8	25.3	**40.6**	27.4	44.3	25.0	**40.3**
+ DRoPE	36.6	50.6	65.9	39.2	52.5	41.3	28.8	24.2	40.2	27.0	44.1	24.2	39.5
GDN-NoPE-LH	37.0	51.1	**68.0**	**40.1**	52.8	39.9	27.1	25.3	40.4	27.4	43.8	**25.7**	39.9
1B Long													
RoPE	38.7	54.4	69.2	43.8	53.8	44.1	27.5	25.8	40.3	27.0	**45.4**	**27.8**	41.5
RoPE-NoPE-LH	37.9	54.0	69.3	44.6	54.7	45.7	28.1	24.2	40.0	25.8	44.8	26.6	41.3
SWA-RoPE-LH	37.0	54.7	69.6	44.7	54.5	**48.0**	28.5	24.8	40.7	27.6	44.6	27.3	41.8
+ DRoPE	37.1	54.1	68.9	44.0	52.6	46.2	28.8	24.8	40.9	**28.0**	45.0	26.6	41.4
SWA-NoPE-LH	**39.0**	53.4	69.2	44.2	51.9	43.9	**30.5**	23.2	**41.5**	27.8	**45.4**	27.2	41.4
GLA-RoPE-LH	37.7	55.2	69.2	**45.8**	55.5	46.4	29.8	24.8	40.5	26.0	43.9	26.7	41.8
+ DRoPE	35.7	55.0	69.4	45.3	55.5	44.1	28.5	26.3	40.6	27.0	43.4	26.2	41.4
GLA-NoPE-LH	37.4	55.8	69.8	45.5	54.2	44.4	29.2	**26.8**	40.6	27.8	43.3	23.5	41.5
GDN-RoPE-LH	38.0	**56.3**	70.2	**45.8**	**56.3**	46.4	28.8	22.2	39.8	27.2	44.4	26.3	41.8
+ DRoPE	37.9	55.9	**71.2**	**45.8**	**56.3**	47.4	29.8	22.2	40.2	27.8	45.0	25.7	**42.1**
GDN-NoPE-LH	37.3	55.8	69.2	45.1	55.8	44.8	29.8	25.8	40.8	27.0	45.1	25.7	41.9
3B Long													
RoPE	36.6	59.9	71.6	50.4	56.3	46.4	29.2	23.7	41.5	**28.2**	45.8	27.2	43.1
RoPE-NoPE-LH	36.4	61.1	72.5	51.3	57.1	47.6	30.9	23.2	41.5	27.4	46.1	26.1	43.4
SWA-RoPE-LH	37.4	61.3	72.1	52.4	57.2	49.9	30.5	27.3	41.1	27.6	48.1	**27.3**	44.3
+ DRoPE	37.6	60.6	71.8	52.4	56.1	50.6	32.5	**28.8**	41.4	25.4	48.0	26.7	44.3
SWA-NoPE-LH	36.0	60.4	72.5	52.2	57.1	50.3	30.5	19.7	**41.8**	27.2	45.9	24.5	43.2
GLA-RoPE-LH	37.4	61.4	72.9	53.4	58.5	51.0	33.2	27.8	40.3	26.4	46.6	26.5	**44.6**
+ DRoPE	37.1	61.4	72.9	**54.2**	56.8	51.7	31.2	25.8	40.7	26.2	46.5	23.9	44.0
GLA-NoPE-LH	37.1	62.3	72.3	51.5	55.4	48.0	**34.2**	21.7	41.7	25.8	46.0	26.5	43.5
GDN-RoPE-LH	**38.5**	**63.7**	**73.1**	52.1	59.3	**52.6**	31.2	24.8	40.5	23.2	**49.0**	26.5	44.5
+ DRoPE	37.6	63.3	73.0	51.9	**60.9**	52.4	31.2	24.8	40.4	25.4	46.5	26.2	44.5
GDN-NoPE-LH	36.6	62.2	72.5	53.1	56.4	52.4	32.9	27.8	**41.8**	26.8	46.5	24.6	44.5

Table 5: Short-context performance of layer-wise hybrid models after long-context pretraining.

	TQA	LMD	PIQA	Hella	Wino	ARC-e	ARC-c	GPQA	SIQA	OBQA	SG	MMLU	Avg.
376M Long													
RoPE	35.2	35.8	64.3	31.2	51.2	36.7	25.8	24.8	38.5	28.0	43.6	24.2	36.6
RoPE-NoPE-HH	**37.9**	36.4	63.6	31.3	50.2	38.5	**27.1**	**26.3**	37.8	28.0	44.1	**26.1**	37.3
SWA-RoPE-HH	35.2	37.6	63.3	31.4	51.0	37.0	22.4	24.8	37.9	28.0	**45.1**	25.1	36.6
+ DRoPE	35.1	37.1	63.3	31.3	52.0	38.1	23.4	24.8	37.2	27.8	**45.1**	25.2	36.7
SWA-NoPE-HH	35.5	36.9	63.8	31.5	51.7	37.2	26.1	25.8	37.9	27.8	43.8	23.5	36.8
GLA-RoPE-HH	37.0	**43.0**	**65.4**	32.4	51.5	**39.2**	24.4	25.8	38.2	24.6	43.9	25.4	**37.6**
+ DRoPE	36.3	42.6	64.4	32.2	51.6	37.2	23.4	25.3	38.3	26.8	44.7	25.1	37.3
GLA-NoPE-HH	36.3	42.4	63.3	32.6	51.9	38.1	25.1	24.2	39.1	27.4	44.7	25.4	37.5
GDN-RoPE-HH	34.9	41.5	63.9	**32.9**	**52.3**	38.8	23.4	24.8	39.0	28.0	43.1	25.7	37.4
+ DRoPE	36.7	40.3	64.3	32.4	51.5	**39.2**	23.7	25.3	**39.2**	**28.2**	43.7	25.7	37.5
GDN-NoPE-HH	36.1	39.1	64.6	32.6	51.6	38.1	25.1	24.2	38.3	27.6	43.6	**26.1**	37.2
776M Long													
RoPE	37.7	47.9	66.1	38.7	52.6	41.1	26.1	25.8	39.6	27.6	43.1	24.6	39.2
RoPE-NoPE-HH	35.7	47.4	67.7	38.5	53.8	40.9	24.4	25.8	39.0	22.6	44.8	26.2	38.9
SWA-RoPE-HH	36.7	46.6	66.3	38.8	51.3	42.2	27.1	24.8	39.3	27.0	43.6	25.0	39.1
+ DRoPE	37.4	45.8	66.4	38.2	52.6	40.7	27.1	26.3	39.5	26.6	44.3	24.4	39.1
SWA-NoPE-HH	35.8	46.3	67.2	39.1	52.8	43.0	24.1	25.3	40.3	27.6	43.0	24.4	39.1
GLA-RoPE-HH	38.3	49.1	68.2	39.4	53.4	42.3	25.4	25.8	40.3	28.0	44.0	25.4	40.0
+ DRoPE	38.0	48.4	68.7	39.5	52.9	41.8	26.1	25.3	**40.8**	28.0	45.3	**26.8**	40.1
GLA-NoPE-HH	37.0	50.3	68.6	39.8	52.6	42.3	25.8	26.3	39.7	26.6	44.0	26.0	39.9
GDN-RoPE-HH	36.4	48.4	69.2	40.1	53.0	43.6	27.5	26.3	40.7	27.4	**45.7**	24.5	40.2
+ DRoPE	35.7	48.3	**69.6**	39.9	53.3	41.3	**30.5**	**26.8**	40.3	**28.2**	45.2	25.4	**40.4**
GDN-NoPE-HH	**38.9**	**51.6**	67.9	**40.2**	**55.0**	**44.8**	27.8	19.7	39.2	27.6	45.5	24.2	40.2
1B Long													
RoPE	**38.7**	54.4	69.2	43.8	53.8	44.1	27.5	25.8	40.3	27.0	45.4	**27.8**	41.5
RoPE-NoPE-HH	36.8	53.6	69.3	44.8	54.0	43.7	28.5	24.2	39.7	28.0	44.8	25.1	41.1
SWA-RoPE-HH	38.5	51.8	69.8	45.0	53.7	45.7	28.8	25.8	41.6	27.2	44.3	24.7	41.4
+ DRoPE	**38.7**	52.1	70.1	44.4	52.9	43.9	27.5	**26.3**	**41.7**	**28.2**	44.8	24.7	41.3
SWA-NoPE-HH	37.9	54.6	69.2	44.7	**55.2**	43.2	29.2	19.7	39.8	25.0	44.5	27.2	40.8
GLA-RoPE-HH	37.4	54.2	69.5	45.2	52.6	47.1	28.1	20.7	41.1	28.0	44.1	25.2	41.1
+ DRoPE	38.0	54.2	69.9	45.6	53.2	45.7	27.1	23.7	40.5	27.6	44.3	24.8	41.2
GLA-NoPE-HH	36.8	56.2	70.2	45.5	53.9	**47.8**	**30.9**	24.2	41.0	27.8	44.3	26.4	**42.1**
GDN-RoPE-HH	36.7	55.2	**70.7**	**45.9**	54.1	44.3	30.2	25.3	40.9	27.4	45.7	25.8	41.8
+ DRoPE	37.0	55.3	70.1	45.3	53.3	43.0	28.1	**26.3**	41.0	**28.2**	45.6	26.6	41.7
GDN-NoPE-HH	37.6	**56.8**	69.4	45.2	54.1	45.2	30.2	20.7	40.8	27.4	**46.1**	26.0	41.6
3B Long													
RoPE	36.6	59.9	71.6	50.4	56.3	46.4	29.2	23.7	41.5	**28.2**	45.8	27.2	43.1
RoPE-NoPE-HH	35.8	61.1	72.1	52.4	56.9	49.7	30.5	23.2	40.5	22.8	44.1	25.5	42.9
SWA-RoPE-HH	36.6	61.1	72.0	51.8	57.6	50.6	32.5	21.7	41.2	23.0	46.0	**28.6**	43.6
+ DRoPE	37.0	60.6	72.0	52.4	56.8	49.9	30.2	25.8	41.3	24.0	45.0	27.5	43.5
SWA-NoPE-HH	36.4	59.9	72.5	51.9	58.2	50.6	29.5	24.8	41.8	25.2	46.0	25.5	43.5
GLA-RoPE-HH	36.1	61.1	72.9	53.4	56.8	**51.3**	31.9	23.7	41.5	26.4	44.7	26.7	43.9
+ DRoPE	37.3	60.3	72.9	53.4	56.8	50.1	30.9	24.2	41.0	24.4	45.5	26.3	43.6
GLA-NoPE-HH	35.8	61.6	**73.4**	53.1	57.7	49.9	30.5	28.3	**42.3**	25.8	45.5	26.6	44.2
GDN-RoPE-HH	**37.6**	**62.7**	72.0	**53.8**	58.2	50.3	33.2	**28.8**	42.1	26.2	45.8	26.5	**44.8**
+ DRoPE	37.4	62.0	73.1	53.1	57.3	49.0	33.9	27.8	42.0	23.4	**46.6**	24.3	44.2
GDN-NoPE-HH	36.8	62.4	72.3	52.5	**60.6**	51.0	**35.3**	25.8	41.8	26.0	45.9	25.7	44.7

Table 6: Short-context performance of head-wise hybrid models after long-context pretraining.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	>4k	All
376M Short																	
RoPE	25.8	0.0	0.0	0.0	0.0	21.7	17.8	8.0	0.4	0.0	0.0	0.0	5.2	6.8	18.3	0.1	6.1
+ NTK	26.5	9.9	0.6	0.1	0.0	21.8	17.8	7.8	8.9	2.2	0.6	0.0	7.4	8.4	18.5	2.8	8.0
+ NTK + Log	26.5	14.0	1.9	0.9	0.5	21.8	17.8	8.0	7.4	7.9	2.2	0.6	8.8	9.4	18.5	4.4	9.1
RoPE-NoPE-LH	29.9	0.1	0.0	0.0	0.0	32.9	19.5	15.7	0.6	0.2	0.0	0.0	6.0	9.8	24.5	0.1	8.2
+ NTK	**30.6**	12.1	1.3	0.7	0.4	33.0	19.5	15.9	23.2	4.9	0.7	0.0	9.0	13.9	24.7	5.4	11.9
+ NTK + Log	**30.6**	18.7	2.2	0.6	0.3	33.0	19.4	15.9	28.4	18.4	5.7	0.9	10.5	17.4	24.7	9.4	14.5
SWA-4k-NoPE-LH	29.9	18.0	6.6	2.8	2.2	33.1	19.5	15.7	24.4	18.7	13.7	7.9	11.9	19.0	24.6	11.8	16.0
+ Log	29.9	18.4	11.5	6.6	3.3	33.0	19.5	15.7	26.8	26.7	25.9	24.7	14.0	24.6	24.5	18.0	20.2
SWA-RoPE-LH	20.8	6.0	0.5	0.2	0.2	32.1	26.5	13.0	9.4	5.8	4.4	3.9	5.5	13.6	23.1	3.8	10.2
SWA-NoPE-LH	29.1	23.5	19.3	11.0	9.2	31.0	27.2	23.9	20.7	17.1	12.9	12.6	18.4	20.8	27.8	15.8	19.8
+ Log	29.1	**24.2**	22.5	24.9	19.4	31.4	27.2	23.9	23.4	22.0	19.3	19.0	24.0	23.7	27.9	21.8	23.9
GLA-RoPE-LH	24.2	3.8	0.7	0.2	0.0	38.1	22.6	13.5	9.4	8.8	6.7	5.1	5.8	14.9	24.6	4.3	11.1
GLA-NoPE-LH	28.6	17.0	2.0	1.3	0.6	**39.0**	**37.0**	33.8	25.5	11.3	2.6	2.8	9.9	21.7	**34.6**	7.9	16.8
+ Log	28.6	19.0	2.6	1.3	0.7	**39.0**	**37.0**	**33.9**	20.6	9.7	3.0	3.1	10.4	20.9	**34.6**	7.5	16.5
+ wsz=4k	28.5	23.1	16.9	10.2	4.5	38.8	**37.0**	33.8	28.0	15.6	13.0	12.9	16.7	25.6	34.5	15.5	21.9
+ Both	28.6	23.7	**26.5**	**26.0**	**19.9**	38.8	**37.0**	**33.9**	**31.4**	27.3	22.9	19.2	**24.9**	**30.1**	**34.6**	**24.6**	**27.9**
GDN-RoPE-LH	21.2	2.0	0.9	0.5	0.0	36.8	31.9	16.6	13.5	4.2	1.4	0.3	4.9	15.0	26.6	2.9	10.8
GDN-NoPE-LH	22.9	20.4	10.4	6.0	3.6	31.8	34.0	30.3	27.0	11.8	1.7	0.1	12.7	19.5	29.7	10.1	16.7
+ Log	22.9	21.2	14.6	7.9	6.9	31.9	34.0	30.3	28.7	16.3	3.8	0.2	14.7	20.7	29.8	12.5	18.2
+ wsz=4k	22.9	21.5	18.1	15.5	7.6	32.0	34.0	30.3	28.2	28.7	21.7	16.9	17.1	27.4	29.8	19.8	23.1
+ Both	22.9	21.9	22.7	20.8	13.6	31.9	34.0	30.3	29.0	**29.5**	**28.4**	**26.2**	20.4	29.9	29.8	24.0	25.9
776M Short																	
RoPE	34.9	0.5	0.0	0.0	0.0	43.9	**39.8**	26.5	1.6	0.7	0.1	0.0	7.1	16.1	36.3	0.4	12.3
+ NTK	36.8	28.6	2.5	0.3	1.8	43.9	39.7	28.4	18.4	5.5	4.8	2.4	14.0	20.4	**37.2**	8.0	17.8
+ NTK + Log	36.9	30.4	12.3	2.8	1.2	44.1	**39.8**	28.2	22.7	10.6	3.0	2.9	16.7	21.6	**37.2**	10.7	19.6
RoPE-NoPE-LH	32.1	0.3	0.0	0.0	0.0	41.1	34.7	30.9	1.9	0.3	2.3	0.0	6.5	15.9	34.7	0.6	12.0
+ NTK	32.3	17.8	0.5	0.1	0.1	41.2	34.6	31.6	24.5	5.9	1.9	0.5	10.2	20.0	34.9	6.4	15.9
+ NTK + Log	32.3	26.1	5.0	1.5	0.3	41.0	34.7	31.3	29.3	20.7	6.9	4.8	13.0	24.1	34.8	11.8	19.5
SWA-4k-NoPE-LH	32.2	22.7	15.5	3.0	1.4	41.0	34.7	30.8	31.0	27.2	21.9	18.5	15.0	29.3	34.7	17.7	23.3
+ Log	32.2	24.2	9.7	7.0	2.5	41.1	34.6	30.8	30.6	**29.3**	25.6	23.3	15.1	**30.8**	34.7	19.0	24.2
SWA-RoPE-LH	22.9	11.4	3.2	1.1	0.8	**46.3**	31.4	20.5	16.8	12.2	9.3	7.5	7.9	20.6	30.3	7.8	15.3
SWA-NoPE-LH	34.9	29.4	23.2	11.7	9.9	35.3	33.6	30.5	28.2	23.4	20.5	16.8	21.8	26.9	33.6	20.4	24.8
+ Log	34.8	32.5	31.7	26.6	25.2	35.3	33.7	30.5	30.6	26.9	24.7	**24.2**	30.2	29.4	33.6	27.8	29.7
GLA-RoPE-LH	31.6	6.7	1.5	1.1	0.5	**46.3**	34.9	24.7	19.5	13.7	11.0	3.0	8.3	21.9	34.4	7.1	16.2
GLA-NoPE-LH	35.0	28.7	10.5	0.5	0.9	40.4	36.3	33.4	28.7	14.9	2.5	2.6	15.1	22.7	36.3	11.2	19.5
+ Log	35.0	29.2	11.5	0.7	0.6	40.3	36.3	**33.5**	31.2	17.6	2.4	2.5	15.4	23.4	36.3	12.0	20.1
+ wsz=4k	35.2	26.9	15.2	11.2	10.9	40.4	36.3	33.4	26.7	19.0	14.8	12.8	19.9	26.2	36.3	17.2	23.6
+ Both	35.1	32.1	30.5	26.1	22.3	40.3	36.3	33.4	**31.3**	28.4	24.3	20.8	29.2	30.7	36.3	27.0	30.1
GDN-RoPE-LH	31.2	9.1	2.0	1.5	1.1	42.1	32.7	22.9	16.1	13.2	10.2	2.1	9.0	19.9	32.2	6.9	15.3
GDN-NoPE-LH	**37.3**	33.4	15.5	11.6	8.8	38.2	32.0	32.1	22.2	12.4	9.5	3.9	21.4	21.5	34.9	14.7	21.4
+ Log	**37.3**	33.7	16.7	11.1	8.6	38.3	32.0	32.0	21.9	10.1	8.0	6.7	21.5	21.3	34.9	14.6	21.4
+ wsz=4k	**37.3**	34.4	34.3	30.4	21.2	38.2	32.0	32.0	27.2	21.9	17.4	15.6	31.5	26.3	34.9	25.3	28.5
+ Both	37.2	**34.7**	**34.6**	**33.0**	**28.3**	38.3	32.0	32.1	29.3	26.8	**25.7**	22.0	**33.6**	29.5	34.9	**29.3**	**31.2**

Table 7: Long-context performance of 376M and 776M LH models after short-context pretraining.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	>4k	All
1B Short																	
RoPE	37.3	1.5	0.0	0.0	0.0	47.4	41.7	28.8	1.7	0.3	0.2	0.1	7.8	17.2	38.8	0.5	13.3
+ NTK	39.3	29.7	6.9	0.5	0.7	47.5	41.6	28.8	21.5	9.0	7.0	2.0	15.4	22.5	39.3	9.7	19.5
+ NTK + Log	39.3	33.4	18.7	8.6	5.8	47.4	41.7	28.8	24.7	20.0	13.4	9.5	21.2	26.5	39.3	16.8	24.3
RoPE-NoPE-LH	46.9	0.1	0.2	0.0	0.0	65.2	**51.9**	**48.1**	2.5	0.1	0.0	0.0	9.4	24.0	53.0	0.4	17.9
+ NTK	45.8	35.9	6.9	0.3	0.2	65.1	51.7	47.0	39.3	13.3	4.2	0.2	17.8	31.5	52.4	12.5	25.8
+ NTK + Log	45.8	37.5	23.7	5.2	1.4	**65.3**	**51.9**	47.0	**45.4**	**41.9**	30.8	9.9	22.7	41.7	52.5	24.5	33.8
SWA-4k-NoPE-LH	46.9	40.4	39.5	34.1	22.2	**65.3**	**51.9**	**48.1**	39.0	31.8	24.9	20.0	36.6	40.1	**53.1**	31.5	38.7
+ Log	46.9	41.1	41.3	42.0	**38.0**	**65.3**	**51.9**	**48.1**	40.5	38.5	**34.3**	28.7	41.9	**43.9**	**53.1**	**38.1**	**43.1**
SWA-RoPE-LH	34.8	14.2	3.3	1.1	0.9	63.3	42.8	28.1	21.3	9.0	8.3	9.6	10.9	26.1	42.2	8.5	19.7
SWA-NoPE-LH	44.8	38.1	29.5	13.0	9.9	53.9	40.7	37.0	31.4	29.8	21.9	17.1	27.1	33.1	44.1	23.8	30.6
+ Log	44.8	**42.6**	**43.0**	40.9	35.1	53.8	40.7	37.0	34.5	34.0	29.0	27.0	41.3	36.6	44.1	35.8	38.5
GLA-RoPE-LH	37.1	12.3	0.5	0.5	0.2	47.7	47.5	31.6	19.4	5.8	6.2	4.2	10.1	23.2	41.0	6.1	17.8
GLA-NoPE-LH	47.0	40.3	13.5	3.2	1.5	42.7	47.2	45.4	38.9	20.3	9.8	8.7	21.1	30.4	45.6	17.0	26.6
+ Log	**47.1**	41.2	20.7	4.3	2.3	42.8	47.3	45.4	40.7	20.1	11.5	10.8	23.1	31.2	45.7	18.9	27.8
+ wsz=4k	47.0	40.2	37.2	35.3	16.8	42.6	47.3	45.4	38.3	30.5	20.9	17.3	35.3	34.6	45.6	29.6	34.9
+ Both	47.0	42.1	41.5	**42.5**	37.5	42.8	47.3	45.4	40.3	34.9	31.9	**29.3**	**42.1**	38.8	45.6	37.5	40.2
GDN-RoPE-LH	41.3	16.0	3.8	1.6	0.7	61.3	50.2	32.8	20.7	13.0	3.8	5.7	12.7	26.8	46.4	8.2	20.9
GDN-NoPE-LH	44.7	27.5	13.8	11.6	9.6	38.6	36.8	37.6	31.7	29.2	15.6	14.7	21.4	29.2	39.4	19.2	25.9
+ Log	44.6	31.1	14.8	11.1	9.4	38.6	36.8	37.6	34.0	29.9	16.9	14.6	22.2	29.8	39.4	20.2	26.6
+ wsz=4k	44.7	36.0	36.6	27.8	18.6	38.6	36.8	37.6	34.9	27.9	21.2	16.1	32.7	30.4	39.4	27.4	31.4
+ Both	44.7	38.8	41.4	38.2	33.5	38.6	36.8	37.6	38.1	32.1	28.8	26.4	39.3	34.1	39.4	34.7	36.2
3B Short																	
RoPE	42.4	1.6	0.1	0.0	0.0	56.8	55.7	40.4	5.9	0.7	0.0	0.0	8.8	22.8	48.8	1.0	17.0
+ NTK	43.7	34.6	15.2	1.1	0.7	56.6	56.0	39.8	28.9	15.1	6.4	0.4	19.1	29.0	49.0	12.8	24.9
+ NTK + Log	43.8	37.0	32.1	19.6	6.5	57.1	55.9	39.5	30.7	26.1	20.4	14.7	27.8	34.9	49.1	23.4	31.9
RoPE-NoPE-LH	56.6	20.5	0.1	0.0	0.0	67.4	**58.7**	52.0	29.0	3.2	2.4	0.4	15.5	30.4	58.7	7.0	24.2
+ NTK	57.1	40.0	19.1	2.4	0.8	67.3	**58.7**	50.7	42.6	25.0	7.8	3.1	23.9	36.5	58.5	17.6	31.2
+ NTK + Log	57.0	45.7	34.3	13.3	5.8	66.9	58.6	50.7	45.9	41.4	32.8	16.7	31.2	44.7	58.3	29.5	39.1
SWA-4k-NoPE-LH	56.5	40.2	36.6	33.3	27.1	**67.5**	**58.7**	52.2	45.8	41.1	30.4	20.4	38.7	45.2	58.7	34.4	42.5
+ Log	56.9	40.3	36.2	31.9	25.7	67.4	58.6	52.1	44.7	**43.2**	**38.5**	32.7	38.2	**48.2**	**58.8**	36.7	44.0
SWA-NoPE-LH	51.8	46.9	44.6	35.5	22.9	64.3	56.5	48.9	41.7	36.6	31.0	22.4	40.3	43.1	55.4	35.2	41.9
+ Log	51.9	**48.9**	**48.8**	**45.9**	**44.0**	63.9	56.4	48.8	44.8	43.1	37.2	32.5	**47.9**	46.7	55.2	**43.2**	**47.2**
GLA-RoPE-LH	41.8	19.3	6.6	0.5	0.1	49.1	54.8	46.7	29.1	12.5	0.6	1.2	13.7	27.7	48.1	8.7	21.9
GLA-NoPE-LH	57.7	35.9	16.5	7.7	9.8	61.5	52.4	48.4	35.2	20.9	24.2	14.5	25.5	36.7	55.0	20.6	32.1
+ Log	57.7	35.5	17.2	11.7	10.0	61.3	52.6	48.8	35.9	12.6	27.0	18.0	26.4	36.6	55.1	21.0	32.4
+ wsz=4k	57.7	43.0	39.4	32.8	22.7	61.6	52.6	48.8	39.8	35.0	26.6	19.8	39.1	40.6	55.2	32.4	40.0
+ Both	**58.0**	43.2	40.8	36.1	31.2	60.9	52.3	49.0	40.6	37.8	35.6	32.0	41.9	44.0	55.1	37.2	43.1
GDN-RoPE-LH	49.9	19.9	6.6	2.3	1.5	53.2	52.0	43.8	30.1	15.8	5.9	4.0	16.1	29.3	49.7	10.8	23.8
GDN-NoPE-LH	52.0	39.1	29.5	15.9	14.3	57.0	56.1	**53.2**	40.2	29.4	21.1	18.2	30.1	39.3	54.6	26.0	35.5
+ Log	51.9	39.5	33.0	22.0	15.9	57.1	56.1	**53.2**	41.8	35.2	31.2	24.3	32.5	42.7	54.6	30.4	38.4
+ wsz=4k	51.9	41.2	38.7	35.0	29.2	57.1	55.9	53.0	44.4	35.8	29.5	22.3	39.2	42.6	54.5	34.5	41.2
+ Both	51.9	41.8	40.5	41.9	35.9	57.1	56.1	52.9	**46.3**	40.7	**38.5**	**33.6**	42.4	46.5	54.5	39.9	44.8

Table 8: Long-context performance of 1B and 3B LH models after short-context pretraining.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	¿4k	All
376M Short																	
RoPE	25.8	0.0	0.0	0.0	0.0	21.7	17.8	8.0	0.4	0.0	0.0	0.0	5.2	6.8	18.3	0.1	6.1
+ NTK	26.5	9.9	0.6	0.1	0.0	21.8	17.8	7.8	8.9	2.2	0.6	0.0	7.4	8.4	18.5	2.8	8.0
+ NTK + Log	26.5	14.0	1.9	0.9	0.5	21.8	17.8	8.0	7.4	7.9	2.2	0.6	8.8	9.4	18.5	4.4	9.1
RoPE-NoPE-HH	33.1	0.3	0.0	0.0	0.0	27.1	28.4	27.0	0.3	0.0	0.0	0.0	6.7	11.8	28.9	0.1	9.7
+ NTK	33.2	14.3	0.6	0.2	0.1	27.3	28.4	27.0	24.6	2.9	1.4	0.1	9.7	16.0	29.0	5.5	13.3
+ NTK + Log	33.2	20.3	1.8	0.2	0.3	27.2	28.4	26.8	23.8	5.2	2.4	0.0	11.2	16.3	28.9	6.8	14.1
SWA-4k-NoPE-HH	33.1	14.6	4.7	2.2	1.2	27.2	28.4	27.0	28.5	27.0	21.0	16.4	11.2	25.1	28.9	14.5	19.3
+ Log (EME)	33.1	21.9	15.6	15.8	9.3	27.0	28.4	27.0	27.0	25.3	26.0	22.9	19.1	26.2	28.9	20.5	23.3
SWA-RoPE-HH	20.6	1.9	0.4	0.4	0.5	36.3	32.8	18.7	11.0	7.4	5.2	4.7	4.7	16.6	27.1	3.9	11.7
SWA-NoPE-HH	29.4	30.0	24.9	17.0	10.5	**44.4**	**33.0**	**31.1**	27.1	26.7	17.0	15.2	22.4	27.8	**34.5**	21.1	25.5
+ Log	29.5	31.2	29.6	**28.7**	**29.1**	44.3	**33.0**	**31.1**	28.6	**29.2**	25.0	22.0	29.6	**30.5**	**34.5**	**27.9**	**30.1**
GLA-RoPE-HH	28.1	2.1	0.3	0.0	0.0	28.5	27.0	17.6	11.0	2.4	1.7	0.2	6.1	12.6	25.3	2.2	9.9
GLA-NoPE-HH	31.2	26.5	6.3	1.3	1.3	20.8	25.6	25.1	24.8	19.7	9.6	6.1	13.3	18.8	25.7	12.0	16.5
+ Log	31.2	27.4	8.0	1.0	0.9	20.9	25.6	25.1	26.5	26.9	13.8	7.9	13.7	21.0	25.7	14.0	17.9
+ wsz=4k	31.2	27.9	19.0	12.0	8.8	20.9	25.7	25.1	20.4	16.3	12.6	11.2	19.8	18.9	25.7	16.0	19.3
+ Both	31.2	28.6	27.7	24.6	16.7	20.9	25.7	25.1	21.6	20.4	18.7	18.6	25.7	21.6	25.7	22.1	23.3
GDN-RoPE-HH	19.2	3.9	0.9	0.6	0.2	32.4	22.5	18.5	11.6	9.8	7.3	6.4	5.0	15.5	23.2	5.1	11.1
GDN-NoPE-HH	**33.5**	30.6	18.4	11.7	7.0	38.4	32.8	29.5	24.9	25.2	24.7	14.0	20.2	27.1	33.5	19.6	24.2
+ Log	**33.5**	**31.9**	24.7	16.5	8.9	38.5	32.8	29.5	26.3	26.3	**29.3**	**25.7**	23.1	29.8	33.6	23.7	27.0
+ wsz=4k	**33.5**	30.0	19.2	10.7	7.4	38.6	32.8	29.5	27.5	21.9	16.1	13.5	20.2	25.7	33.6	18.3	23.4
+ Both	**33.5**	31.5	**30.6**	27.8	25.2	38.6	32.8	29.5	**29.2**	27.3	24.9	21.2	**29.7**	29.1	33.6	27.2	29.3
776M Short																	
RoPE	34.9	0.5	0.0	0.0	0.0	43.9	**39.8**	26.5	1.6	0.7	0.1	0.0	7.1	16.1	36.3	0.4	12.3
+ NTK	36.8	28.6	2.5	0.3	1.8	43.9	39.7	28.4	18.4	5.5	4.8	2.4	14.0	20.4	37.2	8.0	17.8
+ NTK + Log	36.9	30.4	12.3	2.8	1.2	44.1	**39.8**	28.2	22.7	10.6	3.0	2.9	16.7	21.6	37.2	10.7	19.6
RoPE-NoPE-HH	40.7	0.2	0.0	0.0	0.0	49.6	38.7	**37.0**	1.5	0.1	0.0	0.0	8.2	18.1	41.5	0.2	14.0
+ NTK	40.7	33.7	5.8	0.8	0.2	**50.3**	38.8	35.9	29.7	5.9	1.4	0.3	16.2	23.2	41.4	9.7	20.3
+ NTK + Log	**40.8**	**34.9**	6.5	1.2	0.2	50.0	38.7	35.9	**33.9**	7.4	3.7	1.4	16.7	24.4	41.3	11.2	21.2
SWA-4k-NoPE-HH	40.7	28.6	15.8	5.2	1.9	49.9	38.7	36.9	27.3	25.0	20.4	16.1	18.4	30.6	**41.6**	17.5	25.5
+ Log (EME)	40.7	31.2	25.6	22.7	8.5	50.0	38.7	36.9	30.6	28.4	27.1	24.7	25.7	**33.8**	**41.6**	24.9	30.4
SWA-RoPE-HH	32.2	6.6	2.6	0.1	0.0	40.2	31.2	20.8	13.8	11.0	1.8	0.1	8.3	17.0	31.1	4.5	13.4
SWA-NoPE-HH	32.1	29.7	27.9	18.0	10.2	39.5	39.6	34.2	30.8	26.8	18.8	17.1	23.6	29.5	36.3	22.4	27.1
+ Log	32.1	30.7	30.4	**28.3**	**26.6**	39.4	39.7	34.1	33.2	**32.6**	**28.8**	23.8	**29.6**	33.1	36.3	**29.3**	**31.6**
GLA-RoPE-HH	31.0	7.3	1.3	0.4	0.2	41.5	34.5	24.9	18.4	6.8	0.8	2.5	8.0	18.5	33.0	4.7	14.1
GLA-NoPE-HH	35.6	27.2	7.9	8.2	1.9	43.1	36.9	36.0	25.6	3.3	3.9	0.3	16.2	21.3	37.9	9.8	19.2
+ Log	35.7	27.6	9.9	8.8	7.1	43.1	37.0	36.0	25.0	3.4	3.6	0.6	17.8	21.2	38.0	10.8	19.8
+ wsz=4k	35.6	30.0	27.1	15.3	10.5	43.1	36.8	36.0	30.2	26.4	21.1	16.9	23.7	30.1	37.9	22.2	27.4
+ Both	35.6	32.3	**32.0**	25.2	17.8	43.1	37.0	36.0	31.7	29.7	25.1	21.8	28.6	32.1	37.9	27.0	30.6
GDN-RoPE-HH	32.4	6.8	2.7	1.6	1.5	45.2	36.1	23.9	18.6	13.1	10.6	10.5	9.0	22.6	34.4	8.2	16.9
GDN-NoPE-HH	37.8	24.7	12.4	9.8	8.6	43.1	39.2	35.5	23.4	9.0	7.5	7.6	18.7	23.6	38.9	12.9	21.6
+ Log	37.8	25.9	12.9	9.7	9.6	43.2	39.1	35.5	22.6	9.4	7.8	8.5	19.2	23.7	38.9	13.3	21.8
+ wsz=4k	37.7	32.3	24.5	11.5	9.4	43.2	39.4	35.3	29.0	26.9	26.0	23.4	23.1	31.9	38.9	22.9	28.2
+ Both	37.8	33.2	31.6	23.4	16.2	43.1	39.1	35.3	28.0	25.9	25.2	**28.0**	28.4	32.1	38.8	26.4	30.6

Table 9: Long-context performance of 376M and 776M HH models after short-context pretraining.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	¿4k	All
1B Short																	
RoPE	37.3	1.5	0.0	0.0	0.0	47.4	41.7	28.8	1.7	0.3	0.2	0.1	7.8	17.2	38.8	0.5	13.3
+ NTK	39.3	29.7	6.9	0.5	0.7	47.5	41.6	28.8	21.5	9.0	7.0	2.0	15.4	22.5	39.3	9.7	19.5
+ NTK + Log	39.3	33.4	18.7	8.6	5.8	47.4	41.7	28.8	24.7	20.0	13.4	9.5	21.2	26.5	39.3	16.8	24.3
RoPE-NoPE-HH	44.4	1.0	0.1	0.0	0.0	**54.9**	39.7	35.9	11.4	0.0	0.0	0.0	9.1	20.3	43.7	1.6	15.6
+ NTK	43.9	33.7	10.1	1.5	0.8	54.8	39.6	35.6	32.4	13.2	3.7	0.0	18.0	25.6	43.5	11.9	22.5
+ NTK + Log	43.9	35.4	13.6	2.6	0.9	54.8	39.7	35.6	34.4	18.3	3.7	0.0	19.3	26.6	43.5	13.6	23.6
SWA-4k-NoPE-HH	44.4	35.5	34.1	28.3	16.8	54.8	39.7	35.8	31.4	26.1	21.6	17.0	31.8	32.3	43.7	26.4	32.1
+ Log (EME)	44.4	36.4	36.7	34.2	32.8	54.7	39.7	35.8	31.7	31.0	28.9	**28.6**	36.9	35.8	43.7	32.5	36.3
SWA-RoPE-HH	42.0	11.2	2.7	0.5	0.3	53.6	37.7	29.7	25.4	8.2	0.5	0.7	11.3	22.3	40.8	6.2	17.7
SWA-NoPE-HH	43.6	41.1	39.1	29.1	18.8	45.2	43.4	40.2	37.7	31.6	23.5	16.5	34.4	34.0	43.1	29.7	34.2
+ Log	43.7	**42.9**	**44.2**	**41.1**	**41.0**	45.3	43.4	40.2	**39.9**	**35.9**	**31.2**	28.3	**42.6**	**37.7**	43.1	**38.1**	**39.8**
GLA-RoPE-HH	36.8	12.9	4.8	1.4	0.3	37.9	35.4	30.1	22.9	16.9	11.7	1.0	11.2	22.3	35.1	9.0	17.7
GLA-NoPE-HH	**46.6**	28.8	3.2	2.7	1.6	54.4	42.3	**42.1**	36.3	2.3	0.7	0.4	16.6	25.5	**46.4**	9.5	21.8
+ Log	**46.6**	28.3	3.3	2.2	2.0	54.5	42.3	**42.1**	36.3	2.4	1.1	1.4	16.5	25.7	**46.4**	9.6	21.9
+ wsz=4k	**46.6**	30.7	28.2	11.6	4.9	54.5	42.3	**42.1**	34.8	28.3	19.9	11.6	24.4	33.4	**46.4**	21.3	29.6
+ Both	**46.6**	32.0	32.6	30.8	23.9	54.5	42.3	**42.1**	36.3	32.2	27.5	24.6	33.2	37.1	**46.4**	30.0	35.5
GDN-RoPE-HH	41.3	7.5	3.2	0.2	0.4	29.0	**46.5**	29.2	19.4	14.0	4.4	7.6	10.5	21.4	36.5	7.1	16.9
GDN-NoPE-HH	44.4	28.2	12.8	10.6	10.2	36.0	43.6	40.7	26.1	4.4	5.8	6.4	21.2	23.3	41.2	13.1	22.4
+ Log	44.4	26.5	12.8	10.6	9.9	36.1	43.6	40.7	26.5	5.6	4.0	2.9	20.8	22.8	41.2	12.4	22.0
+ wsz=4k	44.4	37.1	36.9	20.9	11.9	36.0	43.6	40.7	33.7	26.1	21.7	15.2	30.3	31.0	41.2	25.5	30.7
+ Both	44.4	37.6	38.8	31.6	31.4	36.1	43.6	40.7	35.6	30.6	28.0	23.4	36.8	34.0	41.2	32.1	35.2
3B Short																	
RoPE	42.4	1.6	0.1	0.0	0.0	56.8	55.7	40.4	5.9	0.7	0.0	0.0	8.8	22.8	48.8	1.0	17.0
+ NTK	43.7	34.6	15.2	1.1	0.7	56.6	56.0	39.8	28.9	15.1	6.4	0.4	19.1	29.0	49.0	12.8	24.9
+ NTK + Log	43.8	37.0	32.1	19.6	6.5	57.1	55.9	39.5	30.7	26.1	20.4	14.7	27.8	34.9	49.1	23.4	31.9
RoPE-NoPE-HH	50.9	24.1	1.1	1.0	1.0	68.7	65.2	**57.5**	12.6	2.3	0.0	0.0	15.6	29.5	60.6	5.3	23.7
+ NTK	51.5	43.1	25.3	3.0	2.7	68.9	65.2	56.2	44.0	21.3	7.0	1.4	25.1	37.7	60.5	18.5	32.5
+ NTK + Log	51.5	43.4	28.5	6.1	2.3	69.0	65.1	55.8	45.0	25.4	7.7	2.1	26.4	38.6	60.4	20.1	33.5
SWA-4k-NoPE-HH	50.8	39.1	35.2	33.5	25.0	**69.5**	65.2	**57.5**	46.4	36.8	26.7	21.5	36.7	46.2	**60.8**	33.0	42.3
+ Log (EME)	50.8	39.6	37.0	37.7	34.2	69.2	**65.3**	**57.5**	**48.1**	40.9	38.5	33.6	39.9	**50.4**	60.7	38.7	**46.0**
SWA-NoPE-HH	**53.9**	46.1	41.9	32.7	21.7	46.3	47.6	45.9	41.6	33.7	28.8	20.9	39.3	37.8	48.4	33.4	38.4
+ Log	**53.9**	**48.2**	**49.5**	**48.0**	**45.6**	46.3	47.7	45.9	44.9	**43.2**	**42.7**	**36.1**	**49.0**	43.8	48.5	**44.8**	**46.0**
GLA-RoPE-HH	44.8	15.7	6.3	0.0	0.0	59.6	43.0	38.4	27.4	18.6	2.2	0.9	13.4	27.2	46.5	8.9	21.4
GLA-NoPE-HH	51.4	25.9	4.3	2.7	2.6	59.4	50.0	42.8	16.5	0.6	0.6	0.2	17.4	24.3	50.9	6.7	21.4
+ Log	51.4	26.1	4.2	3.0	2.4	59.5	49.5	42.7	14.6	0.8	0.2	1.0	17.4	24.0	50.8	6.5	21.3
+ wsz=4k	51.4	42.4	32.5	22.8	15.9	59.6	49.6	42.8	39.2	32.7	23.2	16.9	33.0	37.7	50.8	28.2	35.8
+ Both	51.3	42.5	36.6	30.1	28.7	59.7	50.0	42.8	39.4	35.4	29.3	22.0	37.9	39.8	51.0	33.0	39.0
GDN-RoPE-HH	51.6	20.0	7.3	1.9	1.0	54.4	54.0	38.2	28.4	18.3	11.6	8.6	16.4	30.5	49.6	12.2	24.6
GDN-NoPE-HH	51.3	19.1	12.2	8.9	8.2	58.9	56.6	52.9	30.6	0.5	0.6	0.1	19.9	28.6	54.9	10.0	25.0
+ Log	51.1	29.6	11.6	9.2	8.4	59.2	55.8	52.9	32.9	0.9	1.5	1.2	22.0	29.2	54.8	11.9	26.2
+ wsz=4k	51.2	30.5	37.2	18.8	13.9	58.9	56.1	52.8	32.1	12.1	15.2	15.4	30.3	34.7	54.7	21.9	32.8
+ Both	51.3	26.4	34.9	29.4	31.9	58.6	56.0	52.8	33.9	15.4	13.3	11.2	34.8	34.5	54.7	24.6	34.6

Table 10: Long-context performance of 1B and 3B HH models after short-context pretraining.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	¿4k	All
376M Long																	
RoPE	31.8	26.0	24.9	10.6	2.4	30.1	25.0	21.7	19.1	10.8	5.0	2.6	19.2	16.3	27.2	12.7	17.5
RoPE-NoPE-LH	35.0	28.0	27.1	14.0	3.7	**39.7**	30.1	26.2	27.5	30.5	28.2	19.2	21.6	28.8	32.8	22.3	25.8
SWA-RoPE-LH	30.4	28.6	25.8	12.8	6.1	31.1	27.9	24.6	23.3	25.6	15.5	12.6	20.7	22.9	28.5	18.8	22.0
+ DRoPE	28.2	23.6	23.9	15.0	11.0	30.4	25.6	21.8	20.9	20.1	18.2	16.2	20.3	21.9	26.5	18.6	21.2
SWA-NoPE-LH	**35.4**	28.1	26.8	22.2	17.1	35.3	30.6	29.2	28.0	24.4	20.7	17.3	25.9	26.5	32.6	23.1	26.3
GLA-RoPE-LH	32.8	**30.9**	**30.2**	24.6	10.2	33.6	25.4	25.5	24.7	20.2	12.1	7.3	25.7	21.3	29.3	20.0	23.1
+ DRoPE	32.4	29.6	29.3	27.1	13.6	31.5	25.6	22.8	21.4	16.9	8.1	7.3	**26.4**	19.1	28.1	19.2	22.1
GLA-NoPE-LH	27.5	26.2	25.3	26.9	19.5	30.7	33.2	31.8	30.4	30.0	29.4	23.8	25.1	29.9	30.8	26.4	27.9
GDN-RoPE-LH	27.1	28.3	27.9	12.3	11.0	34.8	31.5	30.5	27.2	22.8	16.3	10.7	21.3	24.8	31.0	19.6	23.4
+ DRoPE	31.2	28.8	29.5	18.9	14.0	35.8	**33.7**	**32.2**	**31.8**	28.4	28.1	24.9	24.5	**30.7**	**33.2**	25.6	28.1
GDN-NoPE-LH	28.2	27.7	27.7	**27.3**	**20.7**	32.6	31.9	29.8	29.1	**31.0**	**30.5**	**30.0**	26.3	**30.7**	30.6	**28.0**	**28.9**
776M Long																	
RoPE	39.4	37.6	37.3	35.5	20.5	44.3	40.2	39.8	36.9	32.8	24.8	22.6	34.1	34.5	40.9	31.0	34.3
RoPE-NoPE-LH	37.7	36.5	35.5	31.9	22.7	43.0	38.1	37.1	32.8	31.1	26.1	22.2	32.9	32.9	39.0	29.9	32.9
SWA-RoPE-LH	**46.2**	38.4	32.5	12.6	6.7	38.0	32.7	31.5	27.8	23.1	16.9	13.7	27.3	26.2	37.1	21.5	26.7
+ DRoPE	42.4	33.5	18.9	10.5	6.6	**50.3**	35.9	32.9	26.4	23.8	20.0	16.0	22.4	29.3	40.4	19.5	26.4
SWA-NoPE-LH	40.6	33.4	30.2	21.2	14.4	47.1	33.7	33.4	30.8	27.3	22.5	18.6	27.9	30.5	38.7	24.8	29.4
GLA-RoPE-LH	43.6	37.1	36.0	33.0	14.5	48.0	35.9	33.6	29.9	25.2	21.7	15.2	32.8	29.9	40.3	26.6	31.1
+ DRoPE	36.4	33.6	31.1	30.9	17.9	40.2	34.2	34.0	31.2	31.0	28.6	21.7	30.0	31.6	36.2	28.3	30.9
GLA-NoPE-LH	38.8	37.9	32.9	27.5	23.7	38.9	40.6	37.7	33.2	32.5	28.4	25.9	32.2	33.9	39.0	30.3	33.2
GDN-RoPE-LH	41.1	37.5	36.0	29.9	19.8	39.6	36.9	32.3	28.1	25.9	20.6	15.3	32.9	28.4	37.5	26.6	30.3
+ DRoPE	37.3	35.0	35.4	29.2	29.1	39.2	38.4	35.9	31.9	33.5	30.8	24.8	33.2	33.5	37.7	31.2	33.4
GDN-NoPE-LH	45.1	**41.0**	**40.2**	**37.4**	**31.9**	48.9	**46.4**	**47.3**	**46.8**	**45.7**	**45.3**	**27.5**	**39.1**	**44.0**	**46.9**	**39.5**	**42.0**
1B Long																	
RoPE	39.6	38.4	36.4	30.9	15.2	48.7	45.3	48.5	41.2	39.0	29.0	24.2	32.1	39.4	45.5	31.8	36.4
RoPE-NoPE-LH	48.1	45.2	44.8	38.5	29.0	53.3	43.4	42.0	41.6	41.6	43.5	35.8	41.1	43.0	46.7	40.0	42.2
SWA-RoPE-LH	43.8	40.7	36.7	24.7	13.1	54.9	44.9	43.6	41.6	36.1	24.9	16.7	31.8	37.5	46.8	29.3	35.1
+ DRoPE	40.3	38.3	35.1	27.0	17.4	42.4	39.1	40.5	34.4	33.0	25.6	19.8	31.6	33.5	40.6	28.8	32.7
SWA-NoPE-LH	50.7	47.5	42.9	29.6	15.7	53.8	47.4	42.9	35.5	34.1	27.3	24.2	37.3	37.9	48.7	32.1	37.6
GLA-RoPE-LH	47.0	44.2	41.7	36.2	20.3	40.1	44.2	40.7	39.2	36.3	33.1	20.1	37.9	36.2	43.0	33.9	36.9
+ DRoPE	48.4	46.4	46.1	40.0	31.0	44.6	47.3	44.5	44.7	44.5	40.8	30.7	42.4	42.4	46.2	40.5	42.4
GLA-NoPE-LH	**55.5**	**49.7**	**49.8**	**44.3**	**41.5**	38.8	44.0	42.8	40.6	40.0	38.4	30.8	**48.2**	39.3	45.3	41.9	43.0
GDN-RoPE-LH	52.3	44.9	41.7	38.0	20.6	**58.7**	50.2	50.7	47.4	46.2	39.5	19.7	39.5	44.6	53.0	37.2	42.5
+ DRoPE	50.3	46.8	44.2	39.6	33.3	54.3	**54.9**	**54.2**	**52.3**	**50.8**	**47.4**	**41.3**	42.9	**50.7**	**53.4**	**44.5**	**47.5**
GDN-NoPE-LH	48.7	43.7	43.0	41.8	28.6	40.7	43.9	42.1	39.6	40.2	38.8	36.5	41.2	40.3	43.8	39.0	40.6
3B Long																	
RoPE	48.3	44.3	41.1	32.0	17.5	58.5	52.2	47.6	49.3	47.1	35.5	24.6	36.6	45.0	51.7	36.4	41.5
RoPE-NoPE-LH	**70.4**	**64.5**	58.9	**50.2**	31.8	67.4	52.4	51.1	50.3	48.5	42.4	**44.9**	55.1	51.0	**60.3**	48.9	**52.7**
SWA-RoPE-LH	52.2	45.8	41.8	33.6	17.1	59.4	46.2	45.0	45.4	46.3	29.8	20.6	38.1	41.8	50.7	35.0	40.3
+ DRoPE	49.3	44.9	39.6	31.8	18.4	52.1	45.7	42.9	40.5	39.0	32.2	26.7	36.8	39.9	47.5	34.1	38.6
SWA-NoPE-LH	56.9	51.4	48.2	41.2	32.5	60.4	52.8	52.4	47.6	44.7	33.9	28.4	46.1	45.7	55.6	41.0	45.9
GLA-RoPE-LH	56.1	52.1	48.4	40.5	26.8	59.8	**59.5**	55.9	47.3	45.7	37.8	24.2	44.8	47.2	57.8	40.4	46.2
+ DRoPE	56.5	53.0	51.4	42.3	38.0	51.4	50.4	48.6	45.2	44.3	41.8	34.1	48.2	45.1	51.7	43.8	46.4
GLA-NoPE-LH	67.2	63.0	**59.2**	49.3	**42.5**	**71.2**	52.7	49.5	43.5	44.0	40.5	37.3	**56.2**	48.4	60.2	47.4	51.7
GDN-RoPE-LH	56.8	49.5	45.2	38.9	22.0	48.3	45.2	42.6	40.9	44.0	40.4	28.3	42.5	41.4	48.2	38.6	41.8
+ DRoPE	54.7	49.5	41.9	41.7	35.6	50.0	51.2	44.4	45.9	45.6	45.2	38.4	44.7	45.8	50.1	43.0	45.3
GDN-NoPE-LH	57.6	52.4	51.2	44.3	37.2	50.1	56.4	**57.5**	**56.7**	**58.0**	**55.3**	42.8	48.5	**53.8**	55.4	**49.7**	51.6

Table 11: Long-context performance of layer-wise hybrid models after long-context pretraining.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	¿4k	All
376M Long																	
RoPE	31.8	26.0	24.9	10.6	2.4	30.1	25.0	21.7	19.1	10.8	5.0	2.6	19.2	16.3	27.2	12.7	17.5
RoPE-NoPE-HH	39.0	33.3	26.3	9.8	7.3	35.9	25.7	27.1	25.3	24.2	14.1	5.9	23.1	22.6	31.9	18.3	22.8
SWA-RoPE-HH	24.3	19.3	16.4	9.7	6.3	39.9	33.8	27.9	27.0	25.0	17.6	13.5	15.2	26.4	31.5	16.8	21.7
+ DRoPE	22.8	20.8	16.6	16.4	14.3	**42.6**	30.9	29.1	29.5	27.2	26.2	23.8	18.2	29.9	31.3	21.9	25.0
SWA-NoPE-HH	35.9	**35.7**	**35.1**	33.7	**32.3**	33.3	**34.5**	**34.5**	**32.9**	**31.5**	**30.5**	**26.7**	**34.5**	**32.0**	**34.6**	**32.3**	**33.1**
GLA-RoPE-HH	34.8	32.3	28.2	26.8	15.4	26.2	21.2	19.5	20.5	17.7	13.8	11.3	27.5	18.6	25.4	20.8	22.3
+ DRoPE	31.4	27.9	29.5	29.1	26.3	34.8	24.1	19.7	21.1	19.8	22.5	15.9	28.8	22.6	27.5	24.0	25.2
GLA-NoPE-HH	33.7	27.5	24.7	19.9	13.3	36.4	34.3	32.9	30.8	27.4	27.7	**26.7**	23.9	30.9	34.3	24.8	28.0
GDN-RoPE-HH	16.3	13.4	13.3	10.0	6.8	38.6	24.2	20.1	18.9	11.2	7.9	8.3	12.0	18.5	24.8	11.2	15.8
+ DRoPE	21.2	19.9	17.5	14.3	11.0	38.5	27.5	22.1	22.1	21.2	19.9	12.9	16.8	23.5	27.3	17.4	20.7
GDN-NoPE-HH	**40.3**	35.2	31.9	**35.0**	23.7	34.8	26.7	27.2	26.1	28.2	26.1	23.1	33.2	27.5	32.3	28.7	29.9
776M Long																	
RoPE	39.4	37.6	37.3	35.5	20.5	44.3	40.2	39.8	36.9	32.8	24.8	22.6	34.1	34.5	40.9	31.0	34.3
RoPE-NoPE-HH	38.6	35.5	31.4	28.4	8.8	38.3	39.4	40.7	37.9	38.1	36.4	30.6	28.5	37.3	39.3	30.9	33.7
SWA-RoPE-HH	40.9	36.1	34.8	30.7	14.7	42.7	36.9	35.9	32.5	26.2	20.4	14.9	31.5	29.9	39.1	26.3	30.6
+ DRoPE	35.3	33.4	32.8	32.7	29.1	41.9	36.6	35.8	33.0	32.4	30.6	27.6	32.7	34.0	37.4	31.4	33.4
SWA-NoPE-HH	35.7	32.2	33.2	32.1	**30.7**	43.3	41.7	39.1	35.3	35.6	33.0	30.2	32.8	36.9	40.0	32.8	35.2
GLA-RoPE-HH	40.7	38.1	36.9	31.7	13.5	46.1	42.5	41.8	36.8	33.7	20.6	16.0	32.2	33.9	42.8	28.4	33.2
+ DRoPE	41.5	37.1	36.9	31.2	10.9	54.3	38.6	37.4	34.9	34.7	31.9	11.3	31.5	34.7	43.0	28.6	33.4
GLA-NoPE-HH	42.4	40.5	**39.8**	36.9	22.5	55.8	**45.9**	**45.4**	**39.9**	37.8	34.2	27.6	36.4	**40.9**	**47.4**	34.9	39.1
GDN-RoPE-HH	42.8	39.9	36.6	28.4	12.8	58.2	39.1	37.3	34.4	34.2	22.5	16.3	32.1	34.6	44.4	28.1	33.5
+ DRoPE	36.0	35.9	34.3	28.8	19.2	**62.1**	44.2	41.4	38.3	38.9	32.4	26.6	30.8	40.6	45.9	31.8	36.5
GDN-NoPE-HH	**45.1**	**43.0**	39.5	**38.1**	28.6	45.1	43.0	40.7	37.7	**39.8**	**36.5**	**37.0**	**38.8**	40.0	43.5	**37.5**	**39.5**
1B Long																	
RoPE	39.6	38.4	36.4	30.9	15.2	48.7	45.3	**48.5**	41.2	39.0	29.0	24.2	32.1	39.4	45.5	31.8	36.4
RoPE-NoPE-HH	51.4	46.4	45.1	40.7	36.4	47.2	41.9	42.1	42.3	37.0	34.9	28.4	44.0	39.1	45.7	38.9	41.2
SWA-RoPE-HH	45.1	41.9	41.8	33.3	16.2	58.1	46.2	42.5	40.9	43.9	28.4	19.4	35.6	39.9	48.0	33.2	38.1
+ DRoPE	41.0	37.7	40.3	34.8	32.9	57.2	44.4	43.6	41.7	40.9	39.4	**38.0**	37.3	43.6	46.6	38.2	41.0
SWA-NoPE-HH	46.8	44.8	43.5	42.4	**43.0**	46.4	41.2	40.2	39.0	39.2	36.9	37.6	44.1	40.1	43.6	40.8	41.8
GLA-RoPE-HH	46.9	40.6	38.9	35.2	16.6	55.4	41.8	40.8	36.4	34.4	29.2	18.1	35.7	36.6	46.2	31.2	36.2
+ DRoPE	44.9	44.5	41.0	37.5	34.6	48.0	42.9	41.6	39.9	39.9	37.0	25.7	40.5	39.3	44.4	37.5	39.8
GLA-NoPE-HH	52.1	49.6	**48.1**	42.5	32.3	**60.7**	**48.5**	47.2	**45.5**	**45.6**	39.9	29.7	44.9	**45.3**	**52.1**	**41.7**	**45.2**
GDN-RoPE-HH	53.1	48.0	42.0	38.7	19.2	50.0	43.6	42.0	42.4	36.6	34.8	18.5	40.2	38.3	47.2	35.0	39.1
+ DRoPE	**54.1**	**49.7**	46.8	**42.9**	35.9	44.7	45.2	41.9	42.9	40.7	37.1	28.6	**45.9**	40.2	46.5	40.6	42.5
GDN-NoPE-HH	51.6	47.1	43.7	37.9	31.9	45.3	43.7	44.0	43.0	41.5	**41.0**	34.8	42.4	41.9	46.2	40.1	42.1
3B Long																	
RoPE	48.3	44.3	41.1	32.0	17.5	58.5	52.2	47.6	49.3	47.1	35.5	24.6	36.6	45.0	51.7	36.4	41.5
RoPE-NoPE-HH	57.5	52.7	48.8	44.1	35.1	58.7	**62.8**	**63.3**	**59.2**	57.3	**52.7**	**45.1**	47.6	**57.0**	**60.6**	**49.4**	**53.1**
SWA-RoPE-HH	50.0	46.7	45.8	40.6	21.2	**66.6**	55.3	53.4	50.4	46.5	36.9	27.2	40.9	48.0	56.3	39.4	45.1
+ DRoPE	46.8	46.2	44.4	41.8	39.8	59.1	48.5	48.1	46.9	46.8	43.1	41.1	43.8	47.7	50.6	43.8	46.1
SWA-NoPE-HH	59.2	**56.7**	50.6	**50.5**	**42.6**	51.0	47.2	45.5	43.0	43.4	43.7	39.8	**51.9**	44.8	50.7	46.3	47.8
GLA-RoPE-HH	53.9	46.8	43.7	36.6	22.4	47.0	44.9	42.8	40.7	35.5	37.0	21.7	40.7	38.5	47.2	35.6	39.4
+ DRoPE	53.8	49.5	43.5	41.2	26.1	63.0	44.0	45.2	41.8	41.4	40.1	30.8	42.8	43.8	51.5	39.3	43.4
GLA-NoPE-HH	57.1	55.2	**53.7**	47.4	38.8	60.9	48.9	48.3	45.8	45.8	45.4	34.9	50.4	47.1	53.8	45.9	48.5
GDN-RoPE-HH	**61.1**	54.3	49.2	38.3	20.9	54.7	47.7	48.8	47.8	46.0	39.0	23.3	44.8	43.9	53.1	39.9	44.3
+ DRoPE	57.6	51.0	45.9	41.2	38.7	47.0	49.4	49.9	47.5	50.9	47.1	39.5	46.9	47.3	51.0	45.2	47.2
GDN-NoPE-HH	58.0	53.3	49.5	43.3	31.5	56.8	53.6	49.1	51.4	**58.1**	**52.7**	17.8	47.1	48.5	54.4	44.7	47.9

Table 12: Long-context performance of head-wise hybrid models after long-context pretraining.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	¿4k	All
376M Long																	
SWA-NoPE-LH	35.4	**28.1**	26.8	22.2	17.1	35.3	30.6	29.2	28.0	24.4	20.7	17.3	25.9	26.5	32.6	23.1	26.3
+ wsz=256	35.1	26.4	24.8	21.7	17.7	35.7	31.0	28.7	28.8	25.5	23.5	19.2	25.2	27.5	32.6	23.5	26.5
+ wsz=512	**37.5**	26.5	**27.4**	24.4	20.9	**39.1**	33.4	29.1	30.4	28.0	23.9	20.1	**27.4**	29.1	**34.8**	25.2	28.4
+ wsz=1k	36.8	26.0	26.5	23.3	22.1	37.6	33.7	30.0	32.1	30.7	29.3	21.7	27.0	30.7	34.5	26.5	29.2
+ wsz=2k	36.2	24.4	**27.4**	**25.8**	**22.8**	34.6	**33.8**	**31.3**	**32.8**	**32.2**	**29.9**	**24.7**	27.3	**31.3**	34.0	**27.5**	**29.7**
+ wsz=4k	36.6	26.1	26.1	24.7	21.2	36.8	32.1	31.1	32.3	30.5	27.6	23.5	26.9	30.6	34.1	26.5	29.0
776M Long																	
SWA-NoPE-LH	40.6	33.4	30.2	21.2	14.4	47.1	33.7	33.4	30.8	27.3	22.5	18.6	27.9	30.5	38.7	24.8	29.4
+ wsz=256	41.8	36.1	31.3	24.0	16.5	48.0	34.9	33.4	31.6	29.1	25.7	19.3	29.9	31.7	39.5	26.7	31.0
+ wsz=512	**45.7**	39.0	35.1	25.2	18.9	**48.4**	35.7	34.0	31.0	30.6	27.3	20.0	32.8	32.4	**41.0**	28.4	32.6
+ wsz=1k	41.9	37.7	32.9	25.1	19.9	45.6	**38.2**	**35.9**	33.0	**32.7**	29.2	22.9	31.5	**33.9**	40.4	29.2	32.9
+ wsz=2k	45.1	38.9	35.0	29.4	25.0	46.6	37.4	33.9	**33.2**	32.2	27.7	23.7	34.7	33.5	40.7	30.6	**34.0**
+ wsz=4k	43.4	**39.2**	**35.2**	**30.8**	**27.1**	45.1	35.4	32.1	31.7	31.4	**29.5**	**23.9**	**35.1**	32.7	39.0	**31.1**	33.7

Table 13: Long-context performance of SWA-NoPE-LH under long-context continual pretraining with larger window sizes (wsz). Larger window sizes exhibit better long-context performance, implying the Short-Window Weariness.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	¿4k	All
376M Short																	
SWA-NoPE-LH	29.1	23.5	19.3	11.0	**9.2**	31.0	27.2	23.9	20.7	17.1	12.9	12.6	18.4	20.8	27.8	15.8	19.8
+ wsz=256	**33.2**	**30.2**	20.8	11.3	8.5	**42.3**	35.2	29.6	24.9	23.1	17.2	15.3	**20.8**	26.8	**35.1**	18.9	**24.3**
+ wsz=512	30.6	26.1	20.6	11.6	8.3	40.5	**35.8**	32.4	28.3	24.5	17.8	11.2	19.5	27.2	34.8	18.6	24.0
+ wsz=1k	32.9	25.9	**22.8**	**12.4**	4.8	26.6	33.8	30.7	28.8	**25.7**	**23.0**	**21.4**	19.8	27.1	31.0	**20.6**	24.1
+ wsz=2k	27.1	18.2	10.2	2.4	1.6	39.9	34.1	**34.1**	**30.6**	25.4	19.7	15.2	11.9	**28.4**	33.8	15.4	21.6
+ wsz=4k	29.9	18.0	6.6	2.8	2.2	33.1	19.5	15.7	24.4	18.7	13.7	7.9	11.9	19.0	24.6	11.8	16.0
376M Short + Log																	
SWA-NoPE-LH	29.1	24.2	22.5	24.9	19.4	31.4	27.2	23.9	23.4	22.0	19.3	19.0	24.0	23.7	27.9	21.8	23.9
+ wsz=256	**33.2**	**31.4**	**29.4**	**28.7**	**27.8**	**42.3**	35.2	29.6	27.3	27.3	24.5	21.6	**30.1**	29.7	**35.1**	**27.3**	**29.9**
+ wsz=512	30.6	27.9	26.8	23.0	16.8	40.6	**35.8**	32.4	30.0	29.2	26.5	22.8	25.0	31.0	34.8	25.4	28.5
+ wsz=1k	29.1	26.7	26.0	25.6	21.2	26.9	33.8	30.7	30.7	30.0	**27.4**	**26.3**	25.7	29.4	30.1	26.7	27.9
+ wsz=2k	27.1	19.9	15.0	13.1	6.3	40.0	34.1	**34.1**	**32.7**	**30.4**	27.0	26.0	16.3	**32.0**	33.8	21.3	25.5
+ wsz=4k	29.9	18.4	11.5	6.6	3.3	33.0	19.5	15.7	26.8	26.7	25.9	24.7	14.0	24.6	24.5	18.0	20.2
776M Short																	
SWA-NoPE-LH	**34.9**	**29.4**	23.2	11.7	**9.9**	35.3	33.6	30.5	28.2	23.4	20.5	16.8	21.8	26.9	33.6	20.4	24.8
+ wsz=256	33.4	26.4	23.0	12.1	9.4	**48.5**	**40.2**	34.0	28.9	25.5	21.3	17.6	20.9	**30.9**	**39.0**	20.5	26.7
+ wsz=512	33.4	28.6	**25.0**	**13.9**	**9.9**	40.9	36.0	**35.0**	**33.3**	**27.5**	20.7	18.3	**22.2**	30.2	36.3	**22.2**	**26.9**
+ wsz=1k	32.9	25.9	22.8	12.4	8.4	40.8	36.2	34.7	28.2	23.0	17.9	16.7	20.5	28.2	36.2	19.4	25.0
+ wsz=2k	30.4	24.5	17.6	6.5	3.5	41.3	39.9	34.3	29.9	26.8	**22.4**	17.9	16.5	30.4	36.5	18.6	24.6
+ wsz=4k	32.2	22.7	15.5	3.0	1.4	41.0	34.7	30.8	31.0	27.2	21.9	**18.5**	15.0	29.3	34.7	17.7	23.3
776m Short + Log																	
SWA-NoPE-LH	**34.8**	**32.5**	**31.7**	26.6	25.2	35.3	33.7	30.5	30.6	26.9	24.7	24.2	**30.2**	29.4	33.6	27.8	29.7
+ wsz=256	33.4	29.1	30.3	**29.0**	26.0	**48.6**	**40.2**	33.9	31.9	29.7	27.7	25.4	29.5	33.9	**39.0**	28.6	32.1
+ wsz=512	33.4	29.7	30.4	28.0	**29.0**	40.9	36.0	**35.0**	**35.7**	**34.7**	**33.0**	**31.5**	30.1	**35.3**	36.3	**31.5**	**33.1**
+ wsz=1k	32.9	27.4	28.6	27.6	21.9	40.9	36.2	34.7	31.8	29.5	27.4	24.8	27.7	32.2	36.2	27.4	30.3
+ wsz=2k	30.4	25.5	23.9	17.8	11.7	41.2	39.8	34.1	31.3	31.2	27.5	25.6	21.8	33.0	36.4	24.3	28.3
+ wsz=4k	32.2	24.2	9.7	7.0	2.5	41.1	34.6	30.8	30.6	29.3	25.6	23.3	15.1	30.8	34.7	19.0	24.2

Table 14: Long-context performance of SWA-NoPE-LH under short-context pretraining with larger window sizes (wsz). Shorter window sizes exhibit better length extrapolation, implying the Long-Window Laziness.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	¿4k	All
376M Long																	
SWA-RoPE-LH	30.4	28.6	25.8	12.8	6.1	31.1	27.9	24.6	23.3	25.6	15.5	12.6	20.7	22.9	28.5	18.8	22.0
+ wsz=2k	30.9	30.9	28.0	16.9	7.6	26.1	25.9	24.7	22.3	21.8	17.8	14.1	22.9	21.8	26.9	19.9	22.3
+ LongCE	32.1	29.3	29.2	28.7	13.5	37.7	31.7	29.0	31.8	32.5	18.8	15.3	26.5	28.1	32.6	24.9	27.5
+ Both	33.7	34.2	32.3	26.8	12.8	**38.7**	31.5	28.3	31.2	**35.2**	21.0	18.4	28.0	29.2	33.1	26.5	28.7
SWA-NoPE-LH	35.4	28.1	26.8	22.2	17.1	35.3	30.6	29.2	28.0	24.4	20.7	17.3	25.9	26.5	32.6	23.1	26.3
+ wsz=2k	36.2	24.4	27.4	25.8	22.8	34.6	33.8	31.3	32.8	32.2	**29.9**	**24.7**	27.3	31.3	34.0	27.5	29.7
+ LongCE	41.4	35.7	34.3	34.7	26.4	35.3	**34.6**	**33.2**	**34.2**	31.3	29.2	23.5	34.5	**31.6**	**36.1**	31.2	32.8
+ Both	**43.4**	**38.7**	**37.7**	**35.7**	**31.0**	35.0	32.3	30.8	31.0	28.3	26.5	24.1	**37.3**	29.7	35.4	**31.6**	**32.9**
776M Long																	
SWA-RoPE-LH	46.2	38.4	32.5	12.6	6.7	38.0	32.7	31.5	27.8	23.1	16.9	13.7	27.3	26.2	37.1	21.5	26.7
+ wsz=2k	**48.0**	35.9	36.2	19.2	8.4	44.7	37.8	**36.7**	33.7	25.7	16.0	15.4	29.5	30.0	**41.8**	23.8	29.8
+ LongCE	45.7	38.3	39.6	24.7	12.1	42.0	39.2	33.8	**39.4**	30.3	20.9	17.8	32.1	31.9	40.2	27.9	32.0
+ Both	45.3	34.0	36.3	33.3	13.2	**48.8**	38.0	33.0	35.7	**36.5**	28.9	18.3	32.4	34.2	41.3	29.5	33.5
SWA-NoPE-LH	40.6	33.4	30.2	21.2	14.4	47.1	33.7	33.4	30.8	27.3	22.5	18.6	27.9	30.5	38.7	24.8	29.4
+ wsz=2k	45.1	38.9	35.0	29.4	25.0	46.6	37.4	33.9	33.2	32.2	27.7	23.7	34.7	33.5	40.7	30.6	34.0
+ LongCE	42.2	36.9	32.9	26.7	18.4	48.3	**39.4**	36.2	32.4	27.4	24.3	21.0	31.4	32.7	41.5	27.5	32.2
+ Both	45.3	**42.0**	**40.3**	**39.7**	**35.3**	44.2	37.9	35.8	33.9	34.7	**30.9**	**29.1**	**40.5**	**35.2**	40.8	**35.7**	**37.4**
1B Long																	
SWA-RoPE-LH	43.8	40.7	36.7	24.7	13.1	54.9	44.9	43.6	41.6	36.1	24.9	16.7	31.8	37.5	46.8	29.3	35.1
+ wsz=2k	48.2	44.3	40.1	28.9	14.9	57.6	43.4	42.5	**42.9**	39.3	28.6	17.8	35.3	38.9	47.9	32.1	37.4
+ LongCE	43.6	42.7	43.5	36.4	19.8	49.9	42.9	43.2	38.9	41.7	33.4	17.4	37.2	38.2	44.9	34.2	37.8
+ Both	44.1	37.4	35.7	37.1	19.9	47.6	47.6	**50.1**	34.4	41.2	**41.3**	14.1	34.8	39.5	47.4	32.6	37.5
SWA-NoPE-LH	50.7	47.5	42.9	29.6	15.7	53.8	47.4	42.9	35.5	34.1	27.3	24.2	37.3	37.9	48.7	32.1	37.6
+ wsz=2k	44.9	44.2	42.1	36.7	28.9	**61.5**	**52.5**	46.3	40.8	39.4	33.3	29.4	39.4	**43.3**	**51.3**	36.9	41.7
+ LongCE	**52.7**	**49.8**	44.1	34.1	21.7	56.0	41.6	38.7	35.9	31.3	27.5	23.9	40.5	36.4	47.2	33.5	38.1
+ Both	48.8	45.2	**45.4**	**41.7**	**37.2**	47.7	46.1	45.6	42.0	**41.9**	37.0	**34.3**	**43.7**	42.1	47.1	**40.6**	**42.7**
3B Long																	
SWA-RoPE-LH	52.2	45.8	41.8	33.6	17.1	59.4	46.2	45.0	45.4	46.3	29.8	20.6	38.1	41.8	50.7	35.0	40.3
+ wsz=2k	57.2	45.4	41.8	33.8	15.5	52.4	47.2	48.7	46.6	49.6	33.8	22.1	38.8	42.9	51.4	36.1	41.2
+ LongCE	52.2	54.4	46.0	42.6	22.0	52.2	45.8	49.2	33.6	45.3	41.4	25.5	43.4	41.9	49.9	38.8	42.5
+ Both	58.2	**59.1**	**53.0**	44.5	26.2	60.6	50.1	49.4	37.0	43.2	44.8	27.0	48.2	44.6	54.6	41.9	46.1
SWA-NoPE-LH	56.9	51.4	48.2	41.2	32.5	60.4	52.8	52.4	47.6	44.7	33.9	28.4	46.1	45.7	55.6	41.0	45.9
+ wsz=2k	58.4	51.7	50.7	43.9	41.1	**63.7**	48.8	47.0	42.5	44.4	40.5	34.2	49.2	45.9	54.5	43.6	47.3
+ LongCE	54.0	50.5	48.0	42.2	37.7	60.4	52.1	54.0	52.4	48.1	36.3	30.2	46.5	47.6	55.1	43.2	47.2
+ Both	**58.9**	57.0	52.9	**47.0**	**41.7**	59.7	**55.9**	**59.9**	**56.6**	**55.7**	**50.3**	**49.3**	**51.5**	**55.3**	**58.6**	**51.3**	**53.7**

Table 15: Long-context performance of SWA layer-wise hybrid models after long-context continual pretraining with extended window sizes (wsz) and LongCE enhancement, both relieving the short-context learning trap effectively.

	NIAH-SK1					NIAH-SK2					NIAH-SK3				
	4k	8k	16k	32k	64k	4k	8k	16k	32k	64k	4k	8k	16k	32k	64k
376M Short															
GLA-NoPE-LH	100.0	0.0	0.0	0.0	0.0	100.0	88.0	0.0	0.0	0.0	39.0	35.0	0.0	0.0	0.0
+ Log	100.0	0.0	0.0	0.0	0.0	100.0	94.0	3.0	0.0	0.0	39.0	51.0	0.0	0.0	0.0
+ SWLA	100.0	70.0	79.0	75.0	46.0	100.0	94.0	57.0	9.0	0.0	39.0	34.0	17.0	16.0	0.0
+ SWLA & Log	100.0	70.0	81.0	84.0	84.0	100.0	99.0	97.0	92.0	68.0	39.0	34.0	52.0	70.0	37.0
GLA-NoPE-HH	100.0	100.0	75.0	0.0	0.0	100.0	98.0	0.0	0.0	0.0	96.0	75.0	0.0	0.0	0.0
+ Log	100.0	100.0	92.0	1.0	0.0	100.0	100.0	0.0	0.0	0.0	96.0	84.0	0.0	0.0	0.0
+ SWLA	100.0	100.0	100.0	100.0	93.0	100.0	98.0	54.0	15.0	5.0	96.0	69.0	29.0	6.0	1.0
+ SWLA & Log	100.0	100.0	100.0	100.0	100.0	100.0	99.0	98.0	91.0	49.0	96.0	81.0	89.0	77.0	37.0
GDN-NoPE-LH	99.0	97.0	88.0	67.0	43.0	75.0	28.0	1.0	0.0	0.0	45.0	82.0	13.0	0.0	0.0
+ Log	99.0	98.0	94.0	91.0	84.0	75.0	37.0	16.0	0.0	0.0	45.0	76.0	25.0	0.0	0.0
+ SWLA	99.0	97.0	82.0	62.0	40.0	75.0	57.0	54.0	31.0	12.0	45.0	65.0	61.0	74.0	23.0
+ SWLA & Log	99.0	97.0	90.0	88.0	79.0	75.0	59.0	71.0	54.0	33.0	45.0	63.0	68.0	71.0	22.0
GDN-NoPE-HH	100.0	98.0	96.0	89.0	79.0	99.0	94.0	37.0	16.0	0.0	99.0	97.0	37.0	2.0	2.0
+ Log	100.0	100.0	100.0	100.0	100.0	99.0	99.0	60.0	30.0	1.0	99.0	98.0	57.0	12.0	2.0
+ SWLA	100.0	100.0	100.0	90.0	82.0	99.0	95.0	20.0	0.0	0.0	99.0	88.0	60.0	28.0	2.0
+ SWLA & Log	100.0	100.0	100.0	100.0	100.0	99.0	99.0	91.0	76.0	71.0	99.0	96.0	93.0	91.0	83.0
1B Short															
GLA-NoPE-LH	100.0	98.0	6.0	0.0	0.0	100.0	100.0	48.0	2.0	0.0	91.0	82.0	8.0	0.0	0.0
+ Log	100.0	98.0	23.0	0.0	0.0	100.0	100.0	77.0	5.0	1.0	91.0	89.0	22.0	0.0	0.0
+ SWLA	100.0	100.0	100.0	100.0	100.0	100.0	100.0	82.0	85.0	33.0	91.0	78.0	59.0	52.0	1.0
+ SWLA & Log	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.0	100.0	99.0	91.0	85.0	73.0	73.0	71.0
GLA-NoPE-HH	96.0	24.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	92.0	80.0	0.0	0.0	0.0
+ Log	96.0	22.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	92.0	77.0	0.0	0.0	0.0
+ SWLA	96.0	35.0	20.0	6.0	1.0	100.0	100.0	99.0	10.0	0.0	92.0	85.0	75.0	46.0	0.0
+ SWLA & Log	96.0	47.0	59.0	77.0	64.0	100.0	100.0	100.0	99.0	68.0	92.0	86.0	81.0	80.0	66.0
GDN-NoPE-LH	100.0	100.0	100.0	100.0	100.0	100.0	69.0	0.0	0.0	0.0	76.0	47.0	0.0	0.0	0.0
+ Log	100.0	100.0	100.0	100.0	100.0	100.0	90.0	4.0	0.0	0.0	76.0	56.0	0.0	0.0	0.0
+ SWLA	100.0	100.0	100.0	100.0	100.0	100.0	91.0	93.0	73.0	32.0	76.0	62.0	60.0	44.0	12.0
+ SWLA & Log	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.0	76.0	66.0	76.0	69.0	64.0
GDN-NoPE-HH	100.0	100.0	100.0	100.0	100.0	100.0	97.0	0.0	0.0	0.0	69.0	17.0	0.0	0.0	0.0
+ Log	100.0	100.0	100.0	100.0	100.0	100.0	95.0	0.0	0.0	0.0	69.0	7.0	0.0	0.0	0.0
+ SWLA	100.0	100.0	100.0	100.0	97.0	100.0	100.0	97.0	36.0	5.0	69.0	62.0	55.0	19.0	9.0
+ SWLA & Log	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0	91.0	96.0	69.0	63.0	61.0	56.0	55.0
3B Short															
GLA-NoPE-LH	100.0	46.0	10.0	33.0	62.0	100.0	98.0	54.0	3.0	0.0	100.0	82.0	25.0	0.0	0.0
+ Log	100.0	50.0	12.0	21.0	39.0	100.0	97.0	51.0	34.0	20.0	100.0	85.0	28.0	12.0	7.0
+ SWLA	100.0	74.0	81.0	71.0	69.0	100.0	100.0	100.0	94.0	53.0	100.0	98.0	82.0	75.0	38.0
+ SWLA & Log	100.0	75.0	67.0	56.0	64.0	100.0	100.0	100.0	99.0	95.0	100.0	99.0	94.0	85.0	69.0
GLA-NoPE-HH	88.0	3.0	0.0	0.0	0.0	100.0	93.0	0.0	0.0	0.0	100.0	41.0	0.0	0.0	0.0
+ Log	88.0	13.0	0.0	0.0	0.0	100.0	80.0	0.0	0.0	0.0	100.0	37.0	0.0	0.0	0.0
+ SWLA	88.0	41.0	15.0	1.0	0.0	100.0	100.0	100.0	86.0	59.0	100.0	92.0	62.0	54.0	26.0
+ SWLA & Log	88.0	44.0	20.0	6.0	5.0	100.0	100.0	100.0	100.0	100.0	100.0	95.0	76.0	69.0	73.0
GDN-NoPE-LH	100.0	100.0	100.0	100.0	100.0	100.0	100.0	88.0	4.0	0.0	95.0	64.0	15.0	0.0	0.0
+ Log	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0	43.0	14.0	95.0	65.0	26.0	9.0	0.0
+ SWLA	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	92.0	90.0	95.0	75.0	66.0	56.0	33.0
+ SWLA & Log	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.0	97.0	95.0	80.0	69.0	68.0	66.0
GDN-NoPE-HH	100.0	100.0	100.0	99.0	99.0	100.0	6.0	0.0	0.0	0.0	96.0	60.0	0.0	0.0	0.0
+ Log	100.0	100.0	100.0	100.0	100.0	100.0	74.0	0.0	0.0	0.0	96.0	58.0	0.0	0.0	0.0
+ SWLA	100.0	100.0	100.0	100.0	96.0	100.0	78.0	100.0	22.0	10.0	96.0	55.0	60.0	21.0	4.0
+ SWLA & Log	100.0	100.0	100.0	100.0	100.0	100.0	39.0	100.0	98.0	99.0	96.0	57.0	58.0	42.0	63.0

Table 16: Extrapolation comparison of the GLA/GDN-NoPE hybrid model on the NIAH-SK1, SK2, and SK3 tasks. Applying a sliding-window linear attention and a log scale for NoPE attention achieves the best length extrapolation.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	¿4k	All
376M Short																	
RoPE	25.8	0.0	0.0	0.0	0.0	21.7	17.8	8.0	0.4	0.0	0.0	0.0	5.2	6.8	18.3	0.1	6.1
RoPE-NoPE-LH	**28.3**	0.1	0.0	0.0	0.0	41.2	27.1	24.8	0.4	0.1	0.0	0.0	5.7	13.4	30.3	0.1	10.2
SWA-NoPE-LH	23.2	16.0	**14.0**	**13.2**	**10.5**	31.8	26.1	21.8	15.8	14.7	12.0	10.6	**15.4**	19.0	25.7	**13.4**	17.5
GLA-NoPE-LH	21.0	3.8	1.5	2.3	2.2	**41.5**	**36.4**	**33.5**	**20.9**	**16.3**	**16.4**	**17.0**	6.2	**26.0**	**33.1**	10.1	**17.7**
GDN-NoPE-LH	26.0	**16.5**	5.1	1.5	0.7	35.4	28.7	22.7	18.4	12.5	10.3	9.9	10.0	19.7	28.2	9.4	15.6
376M Short + Log																	
SWA-NoPE-LH	23.2	**17.8**	**19.4**	**17.6**	**16.9**	31.8	26.1	21.8	17.4	**17.9**	16.7	13.6	**19.0**	20.8	25.7	**17.2**	**20.0**
GLA-NoPE-LH	20.9	6.7	1.6	2.0	1.8	**41.5**	**36.4**	**33.5**	**23.4**	16.9	**17.1**	**17.8**	6.6	**26.7**	**33.1**	10.9	18.3
GDN-NoPE-LH	**26.0**	17.5	12.1	2.3	1.7	35.4	28.7	22.7	19.0	12.7	10.8	8.7	11.9	19.7	28.2	10.6	16.5
376M Long																	
RoPE	31.8	26.0	24.9	10.6	2.4	30.1	25.0	21.7	19.1	10.8	5.0	2.6	19.2	16.3	27.2	12.7	17.5
RoPE-NoPE-LH	**39.7**	**33.8**	24.4	18.6	1.1	**43.8**	34.7	32.1	31.0	29.0	15.4	9.1	23.5	27.9	**37.6**	20.3	26.1
SWA-NoPE-LH	32.8	25.2	23.3	20.1	**17.6**	28.5	29.3	25.7	19.7	16.5	9.7	10.1	23.8	19.9	29.1	17.8	21.5
GLA-NoPE-LH	34.7	32.0	**29.2**	19.1	10.3	38.3	**36.8**	**35.0**	**34.3**	**32.5**	**29.5**	**22.1**	25.1	**32.6**	36.2	**26.1**	**29.5**
GDN-NoPE-LH	34.3	28.3	25.5	**20.8**	**17.6**	35.0	30.3	27.6	24.5	23.5	22.4	19.5	**25.3**	26.1	31.8	22.8	25.8
776m Short																	
RoPE	34.9	0.5	0.0	0.0	0.0	43.9	**39.8**	26.5	1.6	0.7	0.1	0.0	7.1	16.1	36.3	0.4	12.3
RoPE-NoPE-LH	35.9	0.5	0.1	0.1	0.1	42.9	32.6	31.7	4.7	1.2	0.3	0.0	7.3	16.2	35.8	0.9	12.5
SWA-NoPE-LH	31.6	28.1	**25.0**	**19.7**	**15.5**	46.9	**39.8**	**34.1**	**29.5**	**25.5**	20.8	18.8	**24.0**	**30.8**	**38.1**	**22.9**	**27.9**
GLA-NoPE-LH	**37.5**	**30.9**	12.0	3.0	2.7	**48.2**	32.2	26.3	20.0	12.5	10.3	10.8	17.2	22.9	36.1	12.8	20.5
GDN-NoPE-LH	34.1	24.7	14.1	9.8	5.4	40.7	38.1	32.5	25.5	22.1	**21.0**	**19.7**	17.6	28.5	36.4	17.8	24.0
776m Short + Log																	
SWA-NoPE-LH	31.5	29.6	**26.9**	**24.7**	**20.9**	47.0	**39.6**	**34.1**	**31.8**	**30.4**	**28.1**	**22.2**	**26.7**	**33.3**	**38.1**	**26.8**	**30.6**
GLA-NoPE-LH	**37.5**	**32.2**	20.0	7.6	4.4	**48.2**	32.3	26.3	19.9	13.2	10.4	7.8	20.3	22.6	36.1	14.4	21.7
GDN-NoPE-LH	34.0	26.4	16.9	13.4	9.9	40.5	38.0	32.6	26.5	23.0	23.7	22.0	20.1	29.5	36.3	20.2	25.6
776m Long																	
RoPE	39.4	37.6	37.3	35.5	20.5	44.3	40.2	**39.8**	36.9	32.8	24.8	22.6	34.1	**34.5**	40.9	31.0	34.3
RoPE-NoPE-LH	38.2	37.0	36.7	33.3	17.4	43.2	38.0	37.4	**37.5**	**33.1**	27.4	19.5	32.5	33.7	39.2	30.2	33.2
SWA-NoPE-LH	33.2	30.8	31.5	23.8	16.9	**46.7**	**40.4**	38.2	32.7	27.9	23.1	22.0	27.2	33.0	39.6	26.1	30.6
GLA-NoPE-LH	**47.0**	**43.0**	**42.4**	**40.0**	**30.0**	43.7	38.9	35.3	30.9	29.9	**29.0**	**22.8**	**40.5**	32.9	**41.2**	**33.5**	**36.1**
GDN-NoPE-LH	37.4	35.5	34.5	27.0	16.3	42.8	38.6	35.3	33.0	28.6	26.9	21.2	30.2	32.3	38.5	27.9	31.4

Table 17: Long-context performance of layer-wise hybrid models under the YOCO-like layout, verifying the Seesaw Effect across different layer layouts.

	RULER					BABILong							Average				
	4k	8k	16k	32k	64k	0k	2k	4k	8k	16k	32k	64k	RU.	BA	≤4k	¿4k	All
376M Short																	
SWA-NoPE-LH	29.1	23.5	19.3	**11.0**	**9.2**	31.0	27.2	23.9	20.7	17.1	12.9	12.6	18.4	20.8	27.8	15.8	19.8
+ base=5000	27.9	19.6	12.7	8.0	7.0	40.7	34.5	32.1	27.5	24.0	15.9	14.2	15.0	27.0	33.8	16.1	22.0
+ base=2500	28.5	16.2	11.4	9.1	7.7	**45.7**	33.8	28.0	24.0	17.2	11.3	14.6	14.6	24.9	34.0	13.9	20.6
+ base=1000	28.7	21.9	15.5	10.9	7.7	32.6	32.0	28.8	25.4	21.2	13.1	11.4	16.9	23.5	30.5	15.9	20.8
+ base=500	25.6	21.1	17.5	10.7	7.3	32.4	35.0	30.6	26.3	20.8	14.8	14.4	16.4	24.9	30.9	16.6	21.4
+ base=250	22.7	18.7	10.4	5.0	4.1	42.0	**36.7**	32.9	**28.5**	21.6	16.5	13.4	12.2	27.4	33.6	14.8	21.0
+ base=100	**32.2**	**29.3**	**23.5**	10.0	8.8	37.0	35.9	**34.2**	28.4	**24.1**	**19.4**	**15.1**	**20.7**	**27.7**	**34.8**	**19.8**	**24.8**
376M Short + Log																	
SWA-NoPE-LH	29.1	24.2	22.5	24.9	19.4	31.4	27.2	23.9	23.4	22.0	19.3	19.0	24.0	23.7	27.9	21.8	23.9
+ base=5000	27.9	21.9	21.7	19.4	16.0	40.8	34.5	32.1	29.6	29.0	24.5	22.8	21.4	30.5	33.8	23.1	26.7
+ base=2500	28.5	21.5	20.8	18.6	15.9	**45.7**	33.8	28.0	27.5	24.6	22.7	20.6	21.0	29.0	34.0	21.5	25.7
+ base=1000	28.7	25.4	25.9	22.4	22.6	32.8	32.0	28.8	25.9	25.0	25.0	22.2	25.0	27.4	30.6	24.3	26.4
+ base=500	25.6	23.7	23.6	20.5	19.4	32.4	35.0	30.6	28.8	29.4	25.2	23.9	22.6	29.3	30.9	24.3	26.5
+ base=250	22.7	22.8	22.8	19.9	15.8	41.8	**36.7**	32.9	30.5	27.5	25.1	24.2	20.8	31.2	33.5	23.6	26.9
+ base=100	**32.1**	**31.2**	**30.6**	**30.6**	**29.1**	37.0	35.9	**34.2**	**30.9**	**30.3**	**27.5**	**26.0**	**30.7**	**31.7**	**34.8**	**29.5**	**31.3**
376M Long																	
SWA-NoPE-LH	**35.4**	28.1	26.8	22.2	**17.1**	35.3	30.6	29.2	28.0	24.4	20.7	17.3	25.9	26.5	32.6	23.1	26.3
+ base=5000	30.8	23.3	18.4	14.3	7.3	40.0	36.8	33.8	31.7	26.9	22.9	15.9	18.8	29.7	35.3	20.1	25.2
+ base=2500	34.6	27.6	25.1	20.0	10.7	39.4	33.9	32.8	29.1	22.5	17.6	13.3	23.6	26.9	35.2	20.7	25.5
+ base=1000	34.9	28.9	25.6	14.1	9.8	33.5	30.3	27.6	25.8	22.4	19.0	13.2	22.7	24.5	31.6	19.8	23.8
+ base=500	28.0	23.1	22.2	18.4	14.9	35.1	39.9	35.9	**33.1**	25.6	20.8	16.1	21.3	29.5	34.7	21.8	26.1
+ base=250	30.2	27.3	26.8	23.3	14.9	**40.4**	35.4	32.7	31.0	27.9	19.8	16.8	24.5	29.1	34.7	23.5	27.2
+ base=100	32.1	**31.8**	**29.1**	**26.3**	16.6	39.8	**40.9**	**37.4**	32.3	**29.0**	**24.5**	**20.2**	**27.2**	**32.0**	**37.6**	**26.2**	**30.0**
776M Short																	
SWA-NoPE-LH	34.9	29.4	23.2	11.7	9.9	35.3	33.6	30.5	28.2	23.4	20.5	16.8	21.8	26.9	33.6	20.4	24.8
+ base=5000	**36.1**	27.4	17.8	9.8	6.9	37.6	39.8	36.3	29.5	23.3	20.1	16.3	19.6	29.0	37.5	18.9	25.1
+ base=2500	32.1	27.7	19.1	10.7	8.4	37.5	37.3	33.4	25.3	**27.1**	21.3	17.7	19.6	28.5	35.1	19.7	24.8
+ base=1000	33.4	29.3	24.1	12.7	9.6	**52.0**	**44.5**	**38.5**	**29.7**	22.7	20.0	**18.1**	21.8	**32.2**	**42.1**	20.8	**27.9**
+ base=500	33.6	29.3	**27.3**	**20.5**	10.1	43.1	40.2	32.2	26.9	22.3	18.6	17.4	**24.2**	28.7	37.3	**21.6**	26.8
+ base=250	33.7	29.1	24.2	15.8	**10.4**	42.3	33.6	28.7	23.6	20.4	18.8	15.5	22.6	26.1	34.6	19.7	24.7
+ base=100	33.1	**29.6**	22.1	11.1	9.2	42.0	35.0	32.0	27.2	26.2	**21.5**	16.6	21.0	28.6	35.5	20.4	25.5
776m Short + Log																	
SWA-NoPE-LH	34.8	**32.5**	31.7	26.6	25.2	35.3	33.7	30.5	30.6	26.9	24.7	24.2	30.2	29.4	33.6	27.8	29.7
+ base=5000	**36.1**	31.7	**32.8**	26.3	25.6	37.7	39.9	36.3	**33.0**	**31.0**	28.0	**25.4**	30.5	33.0	37.5	**29.2**	32.0
+ base=2500	32.2	30.5	31.1	27.9	26.4	37.3	37.2	33.4	27.7	28.7	23.1	20.4	29.6	29.7	35.0	27.0	29.7
+ base=1000	33.5	30.9	29.7	27.6	27.3	**52.0**	**44.8**	**38.5**	32.4	29.5	**28.2**	25.1	29.8	**35.8**	**42.2**	28.8	**33.3**
+ base=500	33.7	30.1	32.5	**29.7**	27.7	43.0	40.2	32.2	31.2	25.8	24.9	22.8	**30.8**	31.4	37.3	28.1	31.2
+ base=250	33.7	31.7	31.9	27.3	26.8	42.4	33.5	28.7	25.6	22.2	21.5	19.7	30.3	27.7	34.6	25.8	28.8
+ base=100	33.1	32.2	30.3	28.3	**28.4**	41.8	35.0	31.8	28.6	29.9	28.0	24.1	30.5	31.3	35.4	28.7	31.0
776M Long																	
SWA-NoPE-LH	40.6	33.4	30.2	21.2	14.4	47.1	33.7	33.4	30.8	27.3	22.5	18.6	27.9	30.5	38.7	24.8	29.4
+ base=5000	42.4	32.2	29.4	18.6	11.7	42.4	**39.1**	**36.8**	31.1	**29.3**	22.8	18.0	26.9	31.4	40.2	24.1	29.5
+ base=2500	**44.6**	**38.4**	31.9	19.4	13.1	46.8	37.4	33.1	30.9	28.6	23.0	20.0	29.5	31.4	40.5	25.7	30.6
+ base=1000	39.2	36.2	29.3	25.7	19.2	46.7	38.0	36.2	**34.3**	29.2	22.9	**21.0**	29.9	**32.6**	40.0	**27.2**	**31.5**
+ base=500	39.4	36.1	32.1	**29.7**	**23.0**	**49.6**	36.8	36.7	28.0	25.3	21.2	19.0	**32.1**	30.9	**40.6**	26.8	31.4
+ base=250	44.5	38.1	**32.9**	19.4	12.5	46.0	35.7	33.5	27.0	25.6	21.6	19.2	29.5	29.8	39.9	24.6	29.7
+ base=100	42.0	38.2	28.9	21.5	11.9	34.7	30.9	28.9	27.8	26.4	**23.9**	19.9	28.5	27.5	34.1	24.8	27.9

Table 18: Long-context performance of SWA-NoPE-LH with smaller rotary bases under direct extrapolation, log-scaled NoPE extrapolation and context extension.