ROPE-FILTER: IDENTIFYING LONG-RANGE DEPENDENCIES FOR LONG-CONTEXT DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-context modeling is a core capability of large language models. Continued pretraining on long-context data is a standard approach for extending model context windows. Effective long-context extension requires data with strong long-range dependencies, but simply increasing the length of training sequences does not reliably improve long-context capabilities. In this work, we introduce the Frequency-band Selection Gain (FSG), a distance-resolved measure that quantifies how individual RoPE frequency bands contribute to distinguishing attended keys from other keys at the same relative distance. Across multiple LLMs, we find that low-frequency bands consistently exhibit dominant FSG at long distances, and that the boundary separating the active low-frequency and high-frequency regions can be approximately predicted from the RoPE configuration and the distance scale. Based on this observation, we propose RoPE-Filter, which exploits this frequency-dependent long-range behavior for long-context data selection. Specifically, RoPE-Filter weakens the low-frequency contributions to attention logits only in long-range interactions and uses the resulting prediction damage relative to the original model to identify documents with stronger long-range dependencies. Experiments across widely used long-context benchmarks show that models trained on data selected by RoPE-Filter achieve better performance than those trained on data selected by existing open-source methods.

1 INTRODUCTION

Long-context language models are increasingly important for tasks that require understanding long documents and reasoning over retrieved evidence. A common approach to extending the context window of a pretrained language model is to continually pretrain it on long sequences. However, many long-context documents do not contain meaningful dependencies across distant positions. In contrast, documents that require distant information for accurate prediction provide stronger training signals for learning to use long contexts and can more effectively improve the long-context capabilities of pretrained models. Although positional scaling methods can enlarge the accessible context window (Chen et al., 2023; Peng et al., 2024; Ding et al., 2024b), strong long-context capability still depends on the training data. Therefore, identifying documents with meaningful long-range dependencies is critical.

Using an LLM as the scoring model is a common practice in long-context data selection (Wu et al., 2025; Chen et al., 2025; Deng et al., 2026; Chen et al., 2024). Existing approaches typically estimate long-range dependence from observable document- or model-level signals. While such signals can be useful for data selection, they do not directly reveal which internal mechanisms enable the model to select and use distant information. This motivates us to ask a more fundamental question: which components of the model are functionally involved in long-range information selection?

RoPE provides a natural lens for studying this question because it directly couples relative token distance with the attention computation (Su et al., 2024). For a pair of tokens separated by distance Δ, the phase difference in each rotary dimension grows with Δ, while the growth rate is determined by the corresponding rotary frequency. As a result, the contribution of each frequency band changes systematically with token distance.

Motivated by this, we define the Frequency-band Selection Gain (FSG), a distance-resolved measure of how strongly each RoPE frequency band contributes to the keys preferred by the model

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.

within a given distance range, rather than simply measuring its absolute logit magnitude. Across different models and attention heads, we find that FSG becomes increasingly concentrated in low-frequency bands at long distances. More importantly, the active low-frequency region exhibits a clear distance-dependent boundary. This boundary closely follows the RoPE block whose phase reaches approximately half a rotation at the corresponding distance, allowing the active region to be predicted directly from the RoPE configuration and the distance scale. We further validate the functional significance of this region through targeted interventions. Attenuating the predicted low-frequency region selectively degrades long-range retrieval while largely preserving short-context behavior, whereas interventions on neighboring or randomly selected high-frequency bands have substantially smaller effects. Together, these results show that the identified frequency region provides a targeted way to probe the model's use of distant information.

This observation suggests a way to assess long-range dependence at the document level. Intuitively, a document with stronger long-range dependencies should suffer a larger relative increase in perplexity when its long-range information pathway is attenuated. Based on this principle, we propose RoPE-Filter, a training-free method for long-context data selection. RoPE-Filter selectively attenuates the predicted low-frequency region in long-range attention interactions and measures the resulting relative increase in perplexity compared with the original model. We use this increase as the document score and prioritize documents with larger scores for long-context continual pretraining.

Our main contributions are summarized as follows:

- We reveal a distance-dependent frequency structure in long-range information selection. Using Frequency-band Selection Gain (FSG), we find that the active RoPE region shifts toward lower frequencies as distance increases, and that its boundary closely follows the half-rotation scale determined by the RoPE configuration and distance. Targeted interventions further confirm the functional importance of this region for long-range retrieval.
- We propose RoPE-Filter, a training-free long-context data selection method. To the best of our knowledge, we are the first to leverage the frequency-band mechanism of RoPE for long-context data selection.
- Experiments across multiple data domains, scoring models, and training model scales show that models trained on data selected by RoPE-Filter achieve stronger long-context performance than those trained on data selected by existing open-source methods.

2 Related Work

2.1 Pretraining Language Models for Extended Contexts

The long-context ability of a pretrained language model is not determined by the training sequence length alone. Position Interpolation, YaRN, LongRoPE, and related approaches modify or extend positional representations before or during long-context training(Chen et al., 2023; Peng et al., 2024; Ding et al., 2024b; Zhu et al., 2024). The continued pretraining recipe matters, including the training budget, sequence length schedule, learning rate schedule, and attention pattern used during context extension(Xiong et al., 2024; Zhao et al., 2024). The quality and composition of the training data determine which dependencies the model learns. Studies on domain reweighting and data mixture optimization show that domains contribute unequally and their effects depend on the mixture proportions(Xie et al., 2023a; Liu et al., 2025; Li et al., 2024). The ratio of long to short data and the document sampling strategy also play a role in how well the model uses long contexts. Document packing and attention across document boundaries determine which dependencies are available during training (Grattafiori et al., 2024; Fu et al., 2024; Gao et al., 2025c;a). All of these factors influence the performance of the resulting long-context model.

2.2 Data Selection for Language Model Pretraining

Existing data selection methods for language model pretraining primarily target general corpus quality. They combine language identification, text extraction, removal of malformed or low-quality content, deduplication, and heuristic rules to reduce noise and repetition in web data (Wenzek et al., 2020; Xie et al., 2023b; Soldaini et al., 2024; Weber et al., 2024; Qiu et al., 2024). Another line

of work uses model-based quality scores to select documents with higher educational value, informativeness, or overall pretraining utility (Penedo et al., 2024; Li et al., 2024; Wettig et al., 2024). Recent curation pipelines further combine multiple quality classifiers and reduce the use of aggressive heuristic filters to retain more unique content while preserving data quality (Su et al., 2025). These approaches improve the general quality and training value of pretraining corpora, but their selection criteria do not directly measure whether distant tokens contribute to next-token prediction.

Long-context data selection focuses on identifying training samples that contain meaningful dependencies across distant positions, with the goal of making context length extension more efficient. Linguistic metrics can measure properties such as coherence, cohesion, and complexity (Liu et al., 2024). However, these metrics may not fully align with the underlying mechanisms of LLMs. Attention-based approaches measure dependencies between tokens or spans using attention scores (Chen et al., 2025; Wu et al., 2025). Because these approaches primarily rely on the absolute magnitude of attention scores, repeated local patterns can produce consistently high scores even when they do not reflect meaningful long-range dependencies. Other approaches select data based on differences in model signals under long and short context windows, using these differences to estimate the contribution of extended context (Chen et al., 2024; Deng et al., 2026).

2.3 RoPE Frequency Mechanisms

RoPE represents relative position by rotating 2D blocks of query and key vectors at different frequencies (Su et al., 2024). Recent studies show that different frequency bands in RoPE serve different functional roles in language models (Barbero et al., 2025; Hong et al., 2024; Wu et al., 2026). However, these studies primarily analyze absolute attention magnitudes and do not examine the relative differences in attention preferences among keys within a local text span or distance range. They do not systematically study RoPE in relation to long-range dependency modeling and long-context data selection.

3 Methodology

We first study how RoPE frequency bands contribute to information selection across different query–key distances. This analysis reveals a distance-dependent low-frequency region whose boundary can be predicted from the RoPE configuration and distance scale. We then use this frequency structure to construct RoPE-Filter, which measures a document's reliance on distant information through targeted long-range intervention and the resulting change in predictive likelihood.

3.1 Background and Notation

Let d denote the attention head dimension. RoPE partitions each query and key vector into $d/2$ 2D blocks indexed by r. The angular frequency of block r is

$$\theta_r = b^{-2r/d}, \tag{1}$$

where b is the RoPE base. Smaller block indices correspond to higher frequencies, while larger block indices correspond to lower frequencies. For a query at position p and a key at position q, let $\Delta = p - q \geq 0$ denote their relative distance. In RoPE block r, their rotated inner product uses the relative rotation

$$R(-\Delta\theta_r) = \begin{bmatrix} \cos(\Delta\theta_r) & \sin(\Delta\theta_r) \\ -\sin(\Delta\theta_r) & \cos(\Delta\theta_r) \end{bmatrix}. \tag{2}$$

The inner product of the rotated query and key vectors in the 2D RoPE block r is

$$\begin{aligned} c_r(p,q) &= q_p^{(r)\mathsf{T}} R(-\Delta\theta_r) k_q^{(r)} \\ &= \begin{bmatrix} q_{p,1}^{(r)} & q_{p,2}^{(r)} \end{bmatrix} \begin{bmatrix} \cos(\Delta\theta_r) & \sin(\Delta\theta_r) \\ -\sin(\Delta\theta_r) & \cos(\Delta\theta_r) \end{bmatrix} \begin{bmatrix} k_{q,1}^{(r)} \\ k_{q,2}^{(r)} \end{bmatrix} \\ &= \left(q_{p,1}^{(r)} k_{q,1}^{(r)} + q_{p,2}^{(r)} k_{q,2}^{(r)} \right) \cos(\Delta\theta_r) \\ &\quad + \left(q_{p,1}^{(r)} k_{q,2}^{(r)} - q_{p,2}^{(r)} k_{q,1}^{(r)} \right) \sin(\Delta\theta_r). \end{aligned} \tag{3}$$

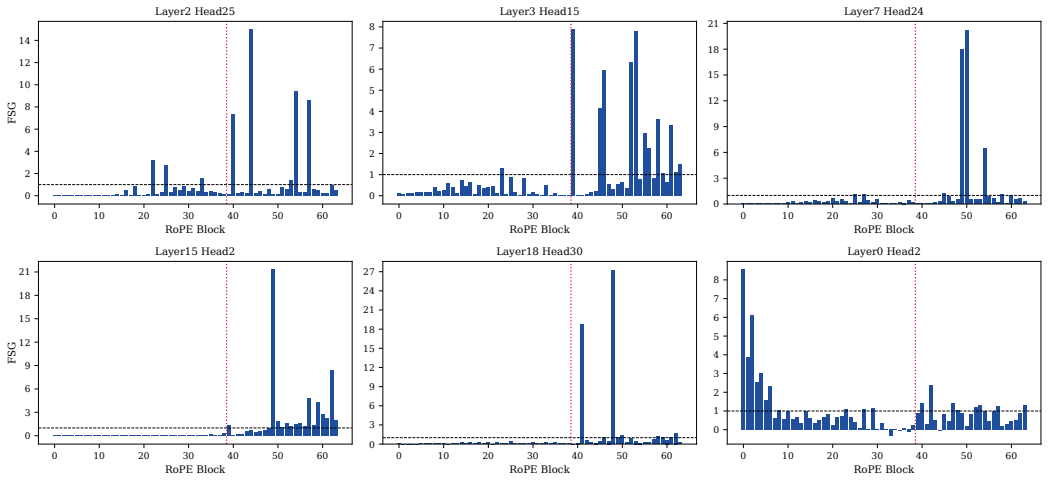


Figure 1: Per-head FSG spectra of Llama-3.1-8B at $\Delta \approx 64$K. Bars are $\hat{s}_r$ per RoPE block. The dashed line marks equal contribution ($\hat{s}_r = 1$) and the dotted line marks the half-rotation boundary $r^* = 39$. The first five heads are representative of the $>90\%$ whose FSG mass concentrates in the low-frequency region $r \geq r^*$. The last panel is a rare exception.

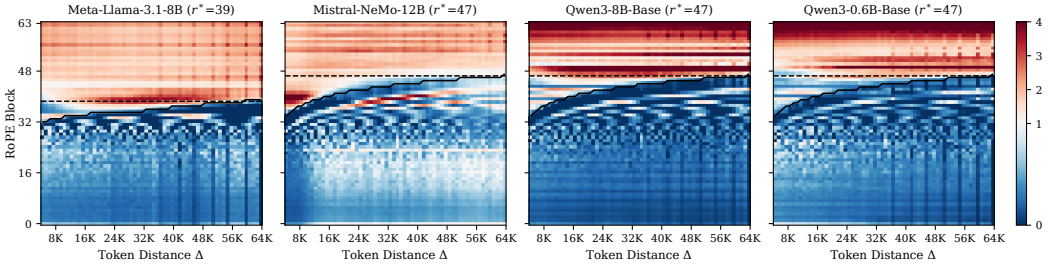


Figure 2: FSG over RoPE blocks and token distances for four models under their native configurations. Color shows $\hat{s}_r(\Delta)$, with 1 denoting equal contribution. The solid line marks the distance-dependent boundary $r^*(\Delta)$, and the dashed line marks the boundary at $T = 64$K.

We write

$$A_r(p,q) = q^{(r)}_{p,1} k^{(r)}_{q,1} + q^{(r)}_{p,2} k^{(r)}_{q,2}, \qquad B_r(p,q) = q^{(r)}_{p,1} k^{(r)}_{q,2} - q^{(r)}_{p,2} k^{(r)}_{q,1}. \tag{4}$$

Then,

$$\begin{aligned} c_r(p,q) &= A_r(p,q)\cos(\Delta\theta_r) + B_r(p,q)\sin(\Delta\theta_r) \\ &= \rho_r(p,q)\cos\big(\Delta\theta_r + \phi_r(p,q)\big), \end{aligned} \tag{5}$$

where

$$\rho_r(p,q) = \sqrt{A_r(p,q)^2 + B_r(p,q)^2} = \left\|q^{(r)}_p\right\| \left\|k^{(r)}_q\right\|, \qquad \tan\big(\phi_r(p,q)\big) = -\frac{B_r(p,q)}{A_r(p,q)}. \tag{6}$$

This representation separates the magnitude and phase offset determined by the query and key vectors from the phase change induced by their relative distance. As the relative distance Δ increases, high-frequency blocks accumulate phase more rapidly, whereas low-frequency blocks vary more slowly and remain more stable over long distances.

3.2 Measuring Block Contributions to Attention

Our intervention targets the RoPE blocks that support the selection of relevant information at long distances. We are motivated by the observation that attention to a retrieved needle forms a distinct

peak relative to surrounding distractors, whereas repetitive data produces a flatter attention distribution with smaller variations among positions within the same distance range. This contrast leads us to measure how RoPE blocks contribute to the model's relative attention preferences among tokens within each distance range.

We first decompose the attention logit into contributions from individual RoPE blocks. Using Eq. 5, the attention logit a_{pq} for a query at position p and a key at position q is

$$a_{pq} = \frac{1}{\sqrt{d}} \sum_{r=0}^{d/2-1} c_r(p,q), \qquad c_r(p,q) = \rho_r(p,q)\cos\big(\Delta\theta_r + \phi_r(p,q)\big), \tag{7}$$

To measure how each RoPE block contributes to the model's attention preference within a given distance range, we define a block-level selection gain. Let π_{pq} be the attention weight from query p to key q, $\mathcal{Q}_p(\Delta)$ the set of keys in a distance bin centered at distance Δ from query p, and $\sigma_p(\Delta)$ the standard deviation of the logits $\{a_{pq} : q \in \mathcal{Q}_p(\Delta)\}$. For block r, we define

$$g_r(\Delta) = \frac{1}{\sqrt{d}\,\sigma_p} \left[\frac{\sum_{q\in\mathcal{Q}_p(\Delta)} \pi_{pq} c_r(p,q)}{\sum_{q\in\mathcal{Q}_p(\Delta)} \pi_{pq}} - \frac{\sum_{q\in\mathcal{Q}_p(\Delta)} c_r(p,q)}{|\mathcal{Q}_p(\Delta)|} \right]. \tag{8}$$

Here, $|\mathcal{Q}_p(\Delta)|$ denotes the number of keys in the distance bin. A larger gain indicates that the block contributes more to keys the model attends to within that bin. To compare the relative contributions of RoPE blocks to attention selection, we define the Frequency-band Selection Gain (FSG):

$$\hat{s}_r(\Delta) = \frac{d}{2} \frac{g_r(\Delta)}{\sum_{r'=0}^{d/2-1} g_{r'}(\Delta)}. \tag{9}$$

To investigate the band patterns exhibited by models when processing long documents, we first probe 500 documents of 64K–128K tokens using 1K-wide distance bins and retain only bins with $\Delta > 4$K. We first examine Llama-3.1-8B at $\Delta = 64$K and find that over 90% of its attention heads exhibit FSG concentrated in low-frequency blocks, as illustrated by the first five panels of Figure 1. Very few heads lack this pattern, as illustrated by the sixth panel. Moreover, the active low-frequency region exhibits a relatively stable boundary.

We observe the same pattern in other models. For example, in Qwen3-8B-Base, low-frequency blocks continue to dominate FSG for bins at $\Delta = 64$K, despite its native 32K context window. This pattern also persists after extending the model to 64K with YaRN without additional training. These observations suggest that low-frequency dominance is shared across model families and is preserved under training-free YaRN extension.

We further probe documents of 8K–64K tokens to examine how this boundary varies with document length. As document length and the query–key distance Δ increase, the boundary of the active low-frequency region shifts toward lower frequencies.

Interestingly, the observed boundary is approximately aligned with the frequency block at which the RoPE phase reaches half a rotation over the corresponding distance scale. Specifically, for a distance T, the half-rotation boundary is given by

$$\mathcal{B}^*(T) = \{r : \theta_r T \le \pi\} = \{r \ge r^*(T)\}, \qquad r^*(T) = \left\lceil \frac{d}{2} \frac{\ln(T/\pi)}{\ln b} \right\rceil, \tag{10}$$

where $\theta_r = b^{-2r/d}$ and b is the RoPE frequency base. To assess how well this boundary matches the observed patterns, we plot FSG across RoPE blocks and query–key distances for four models spanning different sizes and families in Figure 3. At each distance, we aggregate selection gains across all attention heads before normalizing them as in Eq. 9. The solid black curve marks the theoretical boundary $r^*(T)$ with $T = \Delta$, computed from each model's effective RoPE frequencies. Across all four models, this boundary approximately delineates the active low-frequency region. This agreement suggests that RoPE configuration and distance can predict the active low-frequency boundary without model-specific probing.

3.3 Long-range Attention Decay and the Document Score

Following Eq. 10, we select the low-frequency band

$$\mathcal{F} = \mathcal{B}^*(T) = \{r^*(T), \ldots, d/2 - 1\}. \tag{11}$$

For a document X, we define W as scoring window length, $\Delta = p - q$ the query-key distance, D the local window length, and m the number of tokens preserved at the beginning of the document. We define the decay factor as

$$g(p,q) = \begin{cases} 1, & q \le m \text{ or } 0 \le \Delta \le D, \\ 1 - \left(\dfrac{\Delta - D}{W - D}\right)^2, & q > m \text{ and } D < \Delta < W. \end{cases} \tag{12}$$

The decayed attention logit is

$$a_{pq}^{\text{decay}} = \frac{1}{\sqrt{d}} \left[\sum_{r \notin \mathcal{F}} c_r(p,q) + g(p,q) \sum_{r \in \mathcal{F}} c_r(p,q) \right]. \tag{13}$$

This decay weakens the model's use of distant information while preserving the first m keys and the local window D. We evaluate next-token probabilities at positions $\mathcal{T}_D = \{t : D < t \le W\}$. Its perplexity under branch $b \in \{\text{original}, \text{decay}\}$ is

$$\text{PPL}_b(X) = \exp\left(-\frac{1}{|\mathcal{T}_D|} \sum_{t \in \mathcal{T}_D} \log P_b(x_t \mid x_{<t}) \right). \tag{14}$$

We perform one forward pass per branch and score the document by the relative perplexity increase:

$$S(X) = \frac{\text{PPL}_{\text{decay}}(X) - \text{PPL}_{\text{original}}(X)}{\text{PPL}_{\text{original}}(X)}. \tag{15}$$

We retain the documents with the highest scores for continued pretraining. Stronger long-range dependencies are expected to yield larger relative perplexity increases under decay, whereas documents without such dependencies should show little increase. The decay acts on eligible long-range interactions between any pair of positions throughout the scoring window, allowing the score to capture long-range dependencies across the document.

4 Experiments

4.1 Experimental Setup

Long-document Scoring and Selection. From the ArXiv, Book, and CommonCrawl domains of SlimPajama-627B (Soboleva et al., 2023), we use length thresholds of 16K, 64K, and 32K tokens, respectively, to obtain 10B-token long-document pools (ArXiv-Long-10B, Book-Long-10B, and CommonCrawl-Long-10B) and 0.4B-token short-document pools (ArXiv-Short-0.4B, Book-Short-0.4B, and CommonCrawl-Short-0.4B). In large-scale long-context pretraining data selection, scoring long documents at their full lengths is prohibitively expensive. A common practice is to sample the first W tokens to estimate the quality of the entire document. Accordingly, we compare RoPE-Filter with LongAttn (Wu et al., 2025), LADM (Chen et al., 2025), and LongFilter (Deng et al., 2026), using a scoring window of $W = 64$K. We use Qwen3-8B-Base extended to 64K with YaRN as the primary scoring model. Each method selects the 1.6B tokens with the highest scores from each of ArXiv-Long-10B, Book-Long-10B, and CommonCrawl-Long-10B. Additionally, we score documents with Llama-3.1-8B to evaluate the generalizability of our method across different scoring models.

Training Configuration and Evaluation. For each domain, we combine the selected 1.6B long-documents with the same 0.4B short-documents, yielding ArXiv-2B, Book-2B, and CommonCrawl-2B with a long-to-short ratio of 8:2. We pack each dataset into 128K sequences using Best-Fit

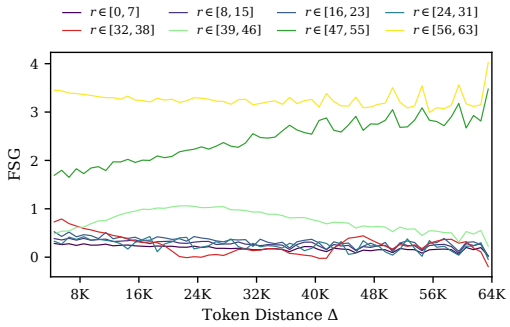


Figure 3: Mean FSG across RoPE blocks within each frequency band as a function of query–key distance in Qwen3-8B-Base with YaRN.

Decreasing (BFD) (Ding et al., 2024a). The main experiments continually pretrain Llama-3-8B-Base on the packed data; the separate larger-model scaling study uses Qwen3-14B-Base. Both use the same optimization recipe, with a model-dependent parallel layout, on 4 nodes with 8 NVIDIA H800 GPUs per node. See Appendix B for the full training configuration and the choice of an 8:2 long-to-short ratio under a 2B token budget. We evaluate the continually pretrained models on common long-context benchmarks to assess the long-range dependencies in the selected data. These benchmarks include HELMET (Yen et al., 2025), RULER (Hsieh et al., 2024), LongBench v2 (Bai et al., 2025), MRCR-v2 (OpenAI, 2025), AA-LCR, and LongProc, covering complementary capabilities in long-context retrieval, multi-hop reasoning, cross-document information synthesis, long-form summarization, and procedural generation. We also evaluate the models on IFEval, BBH, GSM8K, GPQA, and MMLU-Pro to assess whether their short-context capabilities are degraded. For each training configuration, we run CPT and SFT with three random seeds ($\{1, 2, 42\}$) and report the mean across the three trained models. Evaluation details are provided in Appendix C.

RoPE-Filter Setting and Ablations. We use a local window of $D = 4$K and exempt the first $m = 8$ tokens from the decay intervention. Given a scoring model, we apply the decay defined in Eq. 12 to all of its attention heads. Scoring a 10B candidate pool takes approximately 4 hours using 64 NVIDIA H800 GPUs. No further hyperparameter tuning is required. To reduce computational cost, we conduct all ablation and supplementary experiments only on the Book domain, while the main experiments cover all three domains.

4.2 Evaluation across Mainstream Long-Context Benchmarks

In Section 3.2, we find that low-frequency dominance persists after YaRN extension and that Eq. 10 approximately predicts the active region's boundary. For Qwen3-8B-Base extended to 64K with YaRN, applying this criterion to the effective rotary frequencies gives $\mathcal{F} = \{47, \ldots, 63\}$. Figure 3, which aggregates selection gains across all heads at each block and distance, further supports this frequency boundary.

Using Qwen3-8B-Base extended to 64K with YaRN as the scoring model, we compare models trained on data selected by different methods. As shown in Table 1, RoPE-Filter achieves the highest overall score in all three domains. To verify whether the gains in long-context capability come at the cost of short-context performance, we conduct an additional short-context evaluation on the Book domain. The results show that RoPE-Filter improves long-context capability while maintaining strong short-context performance (Appendix A.1).

4.3 Ablation Studies

Scoring Model and Frequency-Band Ablations. The main results use Qwen3-8B-Base with training-free YaRN extension as the scoring model. To further test the robustness of the scoring model choice, we repeat the comparison with Llama-3.1-8B. As shown in Table 2, RoPE-Filter

Table 1: Performance of different data-selection methods using Qwen3-8B-Base extended to 64K with YaRN as the scoring model. Within each domain, the best and second-best results in each column are shown in **bold** and underlined, respectively.

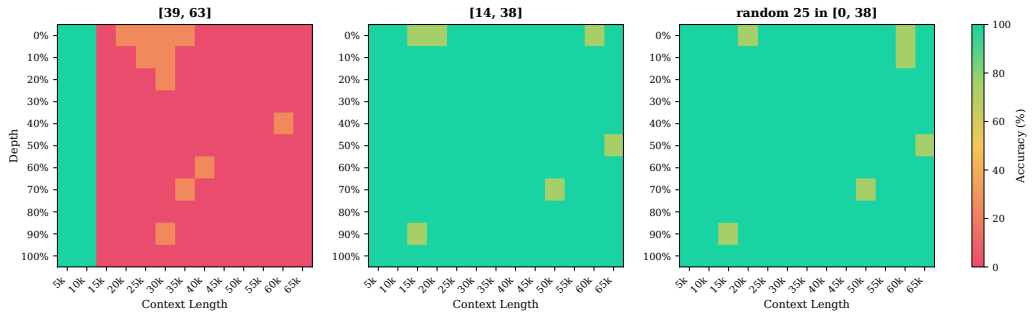
Domain	Method	HELMET	LongBench v2	RULER	MRCR v2	AA-LCR	LongProc	Overall
ArXiv	LongAttn	58.69	**26.50**	74.01	9.78	7.17	9.41	30.92
	LADM	62.71	25.40	72.94	11.42	6.00	**11.71**	31.70
	LongFilter	64.67	25.85	73.27	8.63	6.84	9.54	31.46
	RoPE-Filter	**64.77**	26.10	**74.26**	**12.77**	**8.67**	10.47	**32.84**
Book	LongAttn	64.35	25.95	74.94	18.95	8.00	10.31	33.84
	LADM	63.86	25.90	74.87	16.99	6.33	10.35	33.05
	LongFilter	**65.18**	**27.50**	74.95	17.96	6.84	**11.73**	34.03
	RoPE-Filter	64.71	27.23	**75.66**	**20.26**	**8.33**	10.24	**34.41**
CommonCrawl	LongAttn	61.81	25.80	**76.05**	15.21	5.00	**9.74**	32.27
	LADM	63.25	25.00	74.97	17.64	5.67	8.76	32.55
	LongFilter	**65.29**	26.40	74.57	**18.30**	6.67	8.86	33.35
	RoPE-Filter	64.50	**28.20**	75.56	18.08	**8.67**	8.51	**33.92**
Avg.	LongAttn	61.62	26.08	75.00	14.64	6.72	9.82	32.34
	LADM	63.27	25.43	74.26	15.35	6.00	**10.27**	32.43
	LongFilter	**65.05**	26.58	74.26	14.96	6.78	10.04	32.94
	RoPE-Filter	64.66	**27.18**	**75.16**	**17.04**	**8.56**	9.74	**33.72**

Table 2: Scoring and training model ablations on Book. The first block uses Llama-3.1-8B for scoring and Llama-3-8B for continual pretraining. The second uses Qwen3-8B-Base extended to 64K with YaRN for scoring and Qwen3-14B-Base for continual pretraining.

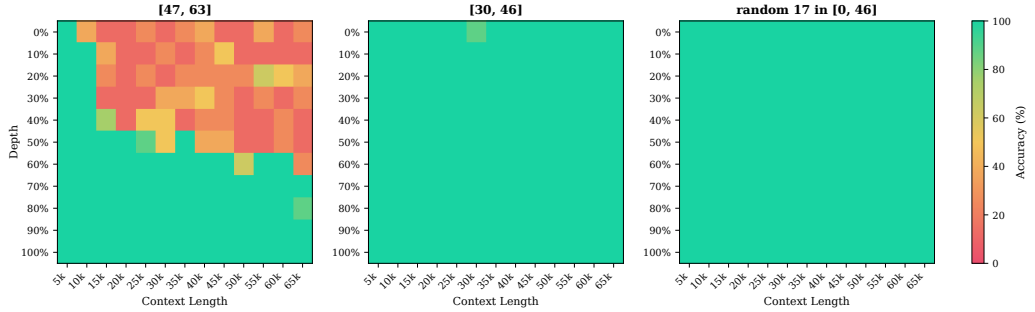
Method	HELMET	LongBench v2	RULER	MRCR v2	AA-LCR	LongProc	Overall
Scoring model ablation							
LongAttn	64.63	25.80	74.07	15.96	7.33	11.07	33.14
LADM	64.65	25.40	75.33	16.32	**8.00**	10.04	33.29
LongFilter	**64.96**	27.43	**75.96**	17.01	7.33	11.22	33.99
RoPE-Filter	64.28	**28.20**	74.58	**20.61**	6.67	**11.50**	**34.31**
Scaling to a larger training model							
LongAttn	65.55	**35.10**	86.62	**22.91**	12.67	21.51	40.73
LADM	65.10	33.00	**86.90**	21.43	13.00	21.47	40.15
LongFilter	64.11	33.80	86.50	21.74	**13.33**	21.66	40.19
RoPE-Filter	**66.14**	34.50	86.52	20.97	12.00	**33.13**	**42.21**

maintains its advantage over the other methods, demonstrating that our method generalizes across different scoring models.

To further explain why it is effective across scoring models, we evaluate the NIAH task on Qwen3-8B-Base extended to 64K with YaRN and Llama-3.1-8B at context lengths from 5K to 65K tokens. For each model, we apply the decay intervention to its corresponding low-frequency RoPE band, $\mathcal{F} = [47, 63]$ for Qwen3-8B-Base and $\mathcal{F} = [39, 63]$ for Llama-3.1-8B. As controls, we decay the same number of blocks selected randomly from the high-frequency region and the same number of blocks immediately adjacent to the low-frequency band, namely $[29, 46]$ and $[14, 39]$, respectively. As shown in Figure 4, decaying the low-frequency blocks impairs long-context retrieval in both models while largely preserving their short-context ability. This result shows that the intervention in Section 3.3 selectively degrades long-context ability. Therefore, the relative PPL increase caused by the intervention measures the strength of a document's long-range dependencies: a larger increase indicates stronger long-range dependence. We further ablate the selected frequency band on the Book domain using Qwen3-8B-Base extended to 64K with YaRN.



(a) Llama-3.1-8B



(b) Qwen3-8B-Base YaRN

Figure 4: NIAH retrieval accuracy under RoPE frequency-band decay. Decaying the predicted low-frequency band impairs long-range retrieval in both scoring models. Decaying an equal number of the nearest higher-frequency blocks or an equal number of randomly sampled higher-frequency blocks has little effect.

Table 3: Downstream results on the Book domain using Qwen3-8B-Base extended to 64K with YaRN as the scoring model. Each row applies RoPE-Filter decay to a different low-frequency band defined by its boundary block.

Block set	HELMET	LongBench v2	RULER	MRCR-v2	AA-LCR	LongProc	Overall
$[45, 63]$	64.74	27.40	75.52	16.21	6.67	11.10	33.61
$[47, 63]^{\star}$	64.71	27.23	75.66	20.26	8.33	10.24	**34.40**
$[50, 63]$	64.37	25.90	76.20	16.89	10.33	11.72	34.24
$[29, 45]$	65.36	27.90	74.77	17.47	7.00	9.22	33.62

Scaling to a Larger Training Model. In practical LLM pretraining, data-selection methods are often first evaluated with a smaller training model and then tested on a larger model to examine whether their benefits scale with model size. Following this setup, we further continually pretrain Qwen3-14B-Base on the data selected by each method. As shown in Table 2, RoPE-Filter yields a larger gain in long-context performance over the strongest baseline on the larger training model. This result suggests that the benefits of RoPE-Filter can scale to larger models. This result highlights the strong potential of RoPE-Filter for large-scale long-document data selection.

5 Conclusion

We introduce FSG to show that low-frequency RoPE bands dominate long-range attention selection, with their active boundary approximately predicted by a phase-based criterion. Based on this finding, RoPE-Filter applies distance-dependent decay to low-frequency contributions for training-free data selection, consistently improving long-context performance across domains and scoring models while preserving competitive short-context performance.

AI USE STATEMENT

Generative AI tools were used to assist with writing and implementation, as well as information retrieval and discovery, including finding related work. Specifically, these tools were used to improve the clarity and readability of text written by the authors, assist with LaTeX formatting, and support routine programming tasks such as code completion and debugging. The research ideas, methodological design, experimental setup, and interpretation of results were developed entirely by the authors without AI assistance. All AI-assisted text and code were reviewed and verified by the authors, who take full responsibility for the final paper, implementation, and reported results.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of RoPE-Filter and the experimental pipeline in the main text and appendices, including the frequency-band selection criterion, long-range attention intervention, document scoring procedure, data construction, sequence packing, training configuration, and evaluation protocols. We compare all data-selection methods using the same candidate pools, selection budgets, short-data mixtures, packing strategy, and training configurations. We also report the relevant model settings, hyperparameters, hardware configuration, random seeds, and benchmark details.

REFERENCES

Artificial Analysis. AA-LCR: Artificial analysis long context reasoning. https://artificialanalysis.ai, 2025.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GtvuNrk58a.

Jianghao Chen, Junhong Wu, Yangyifan Xu, and Jiajun Zhang. LADM: Long-context training data selection with attention-based dependency measurement for LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3076–3090, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.154. URL https://aclanthology.org/2025.acl-long.154/.

Longze Chen, Ziqiang Liu, Wanwei He, Yinhe Zheng, Hao Sun, Yunshui Li, Run Luo, and Min Yang. Long context is not long at all: A prospector of long-dependency data for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8222–8234, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.447. URL https://aclanthology.org/2024.acl-long.447/.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John

Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Haoran Deng, Yingyu Lin, Zhenghao Lin, Xiao Liu, Yizhou Sun, Yian Ma, and Yeyun Gong. Beyond length: Quantifying long-range information for long-context LLM pretraining data. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=C9TDQ8Wwx7.

Hantian Ding, Zijian Wang, Giovanni Paolini, Varun Kumar, Anoop Deoras, Dan Roth, and Stefano Soatto. Fewer truncations improve language modeling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024a.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: extending llm context window beyond 2 million tokens. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Chaochen Gao, Xing W, Qi Fu, and Songlin Hu. Quest: Query-centric data synthesis approach for long-context scaling of large language model. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=sAYnDWaGd5.

Chaochen Gao, Xing W, Zijia Lin, Debing Zhang, and Songlin Hu. NExtlong: Toward effective long-context training without long documents. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=Q9DAI0AYv1.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7376–7399, Vienna, Austria, July 2025c. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.366. URL https://aclanthology.org/2025.acl-long.366/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan

Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj

Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xiangyu Hong, Che Jiang, Biqing Qi, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On the token distance modeling ability of higher RoPE attention dimension. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5877–5888, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.338. URL https://aclanthology.org/2024.findings-emnlp.338/.

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

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dav, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: in search of the next generation of training sets for language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5BjQOUXq7i.

Xiaoran Liu, Kai Lv, Qipeng Guo, Hang Yan, Conghui He, Xipeng Qiu, and Dahua Lin. Long-Wanjuan: Towards systematic measurement for long text quality. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5709–5725, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.327. URL https://aclanthology.org/2024.findings-emnlp.327/.

OpenAI. OpenAI MRCR: Long-context multiple-needle-in-a-haystack benchmark. https://huggingface.co/datasets/openai/mrcr, 2025. OpenAI-MRCR dataset, including the corrected v2 evaluation.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG.

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. URL https://openreview.net/forum?id=wHBfxhZu1u.

Jiantao Qiu, Haijun Lv, Zhenjiang Jin, Rui Wang, Wenchang Ning, Jia Yu, ChaoBin Zhang, Zhenxiang Li, Pei Chu, Yuan Qu, Jin Shi, Lindong Lu, Runyu Peng, Zhiyuan Zeng, Huanze Tang, Zhikai Lei, Jiawei Hong, Keyu Chen, Zhaoye Fei, Ruiliang Xu, Wei Li, Zhongying Tu, Lin Dahua, Yu Qiao, Hang Yan, and Conghui He. Wanjuan-cc: A safe and high-quality open-sourced english webtext dataset, 2024. URL https://arxiv.org/abs/2402.19282.

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. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, 2023. URL https://huggingface.co/datasets/cerebras/SlimPajama-627B.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2459–2475, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.123. URL https://aclanthology.org/2025.acl-long.123/.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and

Ce Zhang. Redpajama: an open dataset for training large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494/.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. Qurating: selecting high-quality data for training language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Longyun Wu, Dawei Zhu, Guangxiang Zhao, Zhuocheng Yu, Junfeng Ran, Xiangyu Wong, Lin Sun, and Sujian Li. LongAttn: Selecting long-context training data via token-level attention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19367–19380, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.991. URL https://aclanthology.org/2025.findings-acl.991/.

Xinyi Wu, Siyuan Liu, and Ali Jadbabaie. How data shapes rope frequency usage: From positional scale matching to length generalization, 2026. URL https://arxiv.org/abs/2607.07678.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: optimizing data mixtures speeds up language model pretraining. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023a. Curran Associates Inc.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023b. Curran Associates Inc.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *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, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.260. URL https://aclanthology.org/2024.naacl-long.260/.

Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. Longproc: Benchmarking long-context language models on long procedural generation. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=ruWC5LIMSo.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context language models effectively and thoroughly, 2025. URL https://arxiv.org/abs/2410.02694.

Liang Zhao, Tianwen Wei, Liang Zeng, Cheng Cheng, Liu Yang, Peng Cheng, Lijie Wang, Chenxia Li, Xuejie Wu, Bo Zhu, Yimeng Gan, Rui Hu, Shuicheng Yan, Han Fang, and Yahui Zhou. Longskywork: A training recipe for efficiently extending context length in large language models, 2024. URL https://arxiv.org/abs/2406.00605.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. PoSE: Efficient context window extension of LLMs via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3Z1gxuAQrA.

A Appendix

A.1 Short-Context Evaluation on the Book Domain

We additionally evaluate short-context capabilities on the Book domain after continued pretraining on data selected by each method. The evaluation covers instruction following, general reasoning, mathematical problem solving, scientific question answering, and broad knowledge understanding. As shown in Table 4, RoPE-Filter achieves the highest aggregate score across the five benchmarks. Although different methods perform best on individual tasks, RoPE-Filter obtains the strongest overall short-context score, indicating that its long-context gains do not require a substantial loss in short-context capabilities.

Table 4: Short-context evaluation on the Book domain. Qwen3-8B-Base extended to 64K with YaRN is used as the scoring model. The best result in the Overall column is shown in **bold**.

Method	IFEval	BBH	GSM8K	GPQA	MMLU-Pro	Overall
LongAttn	38.82	58.14	33.21	29.04	31.99	38.24
LADM	37.34	58.66	34.95	26.77	33.13	38.17
LongFilter	38.63	58.89	34.95	27.78	28.57	37.76
RoPE-Filter	36.97	58.82	35.18	28.54	31.94	**38.29**

B Training Details and Additional Analyses

This section documents the full continued-pretraining (CPT) and instruction-tuning (SFT) setup used in our experiments, together with several ablations that justify our main design choices: the long-to-short data ratio and the training-token budget. Unless otherwise stated, all reported numbers use the setup in Appendix B.6.

B.1 Data Construction and Mixture

We construct the continued-pretraining corpus from three SlimPajama domains: ArXiv, Book, and CommonCrawl. Because the three domains have different length distributions, we use a per-domain length threshold (16K, 64K, and 32K tokens for ArXiv, Book, and CommonCrawl, respectively) to separate long documents from short documents within each domain. For every domain we sample a 10B-token pool of long documents and a fixed set of short documents.

Data selection is applied only to the long documents: each method ranks the 10B-token long-document pool of a domain and keeps the top 1.6B tokens. The short documents are shared across methods and fixed to 0.4B tokens per domain (one short set per domain). The selected long data and the fixed short data are then combined at an 8:2 ratio into a 2B-token per-domain training set (Table 5). This isolates the effect of the selection method, since the only variable across runs is which long documents enter training.

The selected long documents and the fixed short documents are mixed upstream into one per-domain training file before offline packing. The resulting packed cache is then used directly for training.

Table 5: Per-domain data construction. Selection is applied only to long documents; short documents are shared and fixed. The long:short token ratio is 8:2, giving 2B tokens per domain.

Domain	**Long thresh.**	**Long pool**	**Selected long**	**Fixed short**	**Total**
ArXiv	16K	10B	1.6B	0.4B	2.0B
Book	64K	10B	1.6B	0.4B	2.0B
CommonCrawl	32K	10B	1.6B	0.4B	2.0B

B.2 Sequence Packing with Best-Fit Decreasing

Each document is tokenized and wrapped with the model's BOS/EOS tokens, then routed by length. A document longer than the 128K window is cut into consecutive full-window blocks; each such block becomes its own training sequence, and the trailing remainder is kept only if it is at least $10{,}240$ tokens, in which case it re-enters the packing pool (otherwise it is dropped). All documents no longer than the window, together with the recycled remainders, form the packing pool.

We pack the pool into 128K ($131{,}072$-token) sequences using Best-Fit-Decreasing (BFD) packing (Ding et al., 2024a): documents are sorted by length in decreasing order, and each is placed into the bin whose remaining capacity is smallest among those that can still hold it; if none can, a new bin is opened. BFD never splits a document to fill a bin, preferring to leave padding, which yields a high token fill rate (> 0.94 in every domain) while keeping each document intact.

Within a packed sequence, documents are kept mutually independent. We restart the position IDs at zero for every document and apply a block-diagonal attention mask, so each token attends only to earlier tokens *within the same document* and never across a document boundary. This follows the document-masked packing used by ProLong (Gao et al., 2025c) and prevents positional and attention signals from leaking between co-packed documents. To align with context parallelism, each document is padded to a multiple of $2 \times \mathrm{CP} = 8$ tokens with the end-of-sequence token. The loss label of every padding position is set to -100, as is the label of the first token of each document, which masks the boundary loss under the standard causal label shift.

B.3 Run-to-Run Stability (Noise Floor)

Before reading the ablations that follow, we quantify how much of the observed score movement is simply run-to-run noise. We repeat the full Llama-3-8B CPT+SFT pipeline with three different random seeds ($\{1, 2, 42\}$) using the same fixed 1B data subset and the same non-budget hyperparameters as Appendix B.6, and evaluate each run on the six long-context tasks. Across the three seeds the long-context composite spans a range of only 0.17 points (std. 0.09; Table 6), and it is more stable than any of the six benchmarks it averages—aggregating over six tasks cancels most of the per-task noise, so the composite is the quantity we read the ablations on. With only three seeds we treat the range as the more robust of the two statistics.

This noise floor is the yardstick the following ablations are read against. The effects we rely on are an order of magnitude larger than it: the $1 \rightarrow 2$B budget step lifts the long-context composite by $+1.93$ ($11\times$ the range, Appendix B.5) and moving from a pure-long mixture to $8{:}2$ lifts it by $+2.33$ ($14\times$, Appendix B.4), while the margin of RoPE-Filter over the strongest baseline in Table 1 ($+0.69$ long-context average) is still about $4\times$ the range. Conversely, movements of a few tenths—such as each individual post-2B budget step—are only a small multiple of the floor and we do not draw conclusions from their ordering. These stability runs use a fixed data subset at the 1B budget, so their absolute level is not comparable to the ablation tables; we use them only to estimate variance.

Table 6: Run-to-run stability of the Llama-3-8B CPT+SFT pipeline: three repeats under an identical configuration, varying only the random seed. We report the long-context composite only—the mean of the six long-context tasks, matching the convention of Table 1 (MRCR-32K excluded, MRCR-v2 retained). Range is max−min across the three seeds.

	Seed 1	**Seed** 2	**Seed** 42	**Mean**	**Range**	**Std. dev.**
Long-context Avg.	32.74	32.57	32.61	32.64	**0.17**	**0.09**

B.4 Long-to-Short Ratio Ablation

The $8{:}2$ long-to-short mixture is fixed for all runs and is not tuned per method. Mixing a fraction of short documents into a long-context corpus serves two purposes, both emphasized by ProLong (Gao et al., 2025c): it preserves the model's short-context ability and guards against catastrophic forgetting during the long-context continued-pretraining stage, and—because short, high-quality documents remain in the stream—it can further elicit long-context ability rather than merely trading

it away. We therefore sweep the long:short token ratio and evaluate both long- and short-context performance (Table 7).

On our Llama-3-8B base, at the 2B-per-domain budget used throughout, a 20% short admixture (8:2) attains the best overall score (37.81) and the best short-context score (40.44), while its long-context composite (35.18) is within 0.4 points of the best (6:4, 35.53). The pure-long setting (10:0) is the weakest on every aggregate—worst long-context (32.85) and worst overall (36.46)—so mixing in short documents does *not* trade away long-context ability. In fact long-context performance is monotone in the short fraction (10:0 $<$ 8:2 $<$ 6:4), which is exactly ProLong's point that high-quality short documents can further elicit long-context ability rather than merely preserving short-context skill. Pushing the mixture on to 6:4 buys a little more long-context but gives back short-context (and hence overall) performance. The 8:2 mixture therefore sits at the knee of the trade-off—it maximizes short-context retention and overall quality while keeping long-context close to the best (0.35, about twice the noise floor of Appendix B.3), whereas the gap it opens over the pure-long setting ($+2.33$ long-context) is more than an order of magnitude above that floor. We adopt it for all main runs.

Table 7: Effect of the long-to-short token ratio on our Llama-3-8B base under the 128K CPT+SFT pipeline, at the 2B-per-domain continued-pretraining budget used throughout. Long Avg. is the mean of six long-context tasks (HELMET, LongBench v2, RULER, MRCR-v2, AA-LCR, LongProc), matching the convention of Table 1; Short Avg. is the mean of five short-context tasks (IFEval, BBH, GSM8K, GPQA, MMLU-Pro). The best result in each column is in **bold**.

Long:Short	Long Avg.	Short Avg.	Overall
10:0	32.85	40.07	36.46
8:2	35.18	**40.44**	**37.81**
6:4	**35.53**	39.65	37.59

B.5 Training-Budget Ablation

We use a 2B-token per-domain budget throughout. To justify this choice we sweep the continued-pretraining budget from 1B to 5B tokens at the fixed 8:2 mixture, SFT each checkpoint identically, and evaluate long- and short-context performance (Table 8, Figure 5). Because every main-paper model is continued-pretrained on 2B tokens *per domain* (three independent per-domain models, later averaged), the 2B point of this sweep corresponds exactly to the training maturity used in all main runs.

The scaling curve has a clear knee at 2B. At 1B the model is under-trained: the long-context composite is only 33.25, and the differences that the continued-pretraining stage is meant to expose have not yet developed. The single largest improvement in the whole sweep is the $1 \to 2$B step, which lifts the long-context composite by $+1.93$ and the overall score by $+1.40$—an order of magnitude above the noise floor of Appendix B.3. Beyond 2B the curve flattens sharply: no further billion tokens moves the overall score by more than ± 0.45 or the long-context composite by more than ± 0.65 (the $2 \to 3$B step is even slightly negative on long-context, -0.40), and going all the way from 2B to 5B—more than doubling the compute—adds only $+0.41$ overall over what 2B already achieves. Put differently, 2B already captures roughly four-fifths of the long-context gain available anywhere in the sweep while remaining cheap, and it is large enough to separate selection methods without over-training on a fixed long-document pool. We therefore adopt 2B tokens per domain as the best compute/performance trade-off for all main runs.

B.6 Training Configuration

Our main experiments use Llama-3-8B-Base as the training model; the scoring model is separate and does not determine the training recipe. We extend the training context to 128K through continued pretraining, followed by instruction tuning. A separate larger-model scaling study uses Qwen3-14B-Base. Training uses MS-Swift with a Megatron backend. Before continued pretraining we set the RoPE base to $\theta = 2 \times 10^8$ (ABF), following NExtLong (Gao et al., 2025b). Tables 9 and 10 summarize the training settings.

Table 8: Effect of the continued-pretraining token budget at a fixed 8:2 mixture, on our Llama-3-8B base under the 128K CPT+SFT pipeline. Long Avg. is the mean of six long-context tasks (HELMET, LongBench v2, RULER, MRCR-v2, AA-LCR, LongProc), matching the convention of Table 1; Short Avg. is the mean of five short-context tasks (IFEval, BBH, GSM8K, GPQA, MMLU-Pro). The $1 \rightarrow 2$B step gives the single largest jump; gains beyond 2B (our default, in **bold**) diminish.

CPT budget	Long Avg.	Short Avg.	Overall
1B	33.25	39.57	36.41
2B (default)	**35.18**	**40.44**	**37.81**
3B	34.78	40.67	37.73
4B	35.42	40.88	38.15
5B	35.71	40.72	38.22

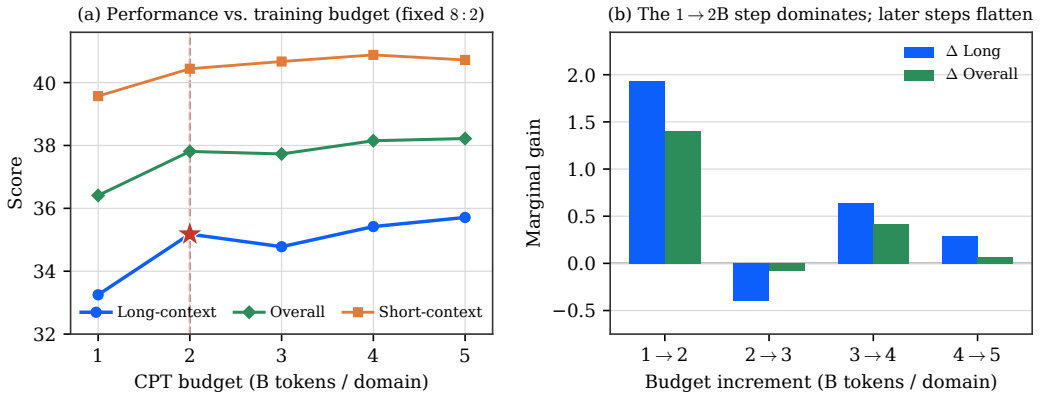


Figure 5: Continued-pretraining token-budget ablation on our Llama-3-8B base at a fixed 8:2 mixture (128K CPT+SFT pipeline; budget is the only variable). **(a)** Long-, short-, and overall-context composites versus the per-domain CPT budget. Performance rises steeply from 1B to 2B and then enters a shallow diminishing-returns regime; our default 2B (red star, dashed line) sits right at the knee. **(b)** Marginal gain per additional 1B tokens: the $1 \rightarrow 2$B step dominates ($+1.93$ long, $+1.40$ overall), while every later step contributes at most ± 0.65 and $2 \rightarrow 3$B is slightly negative on long-context. Exact values are listed in Table 8.

Table 9: Continued-pretraining hyper-parameters for the main Llama-3-8B-Base experiments. The separate Qwen3-14B-Base scaling study uses the same optimization settings but TP/CP/PP/DP $= 4/4/1/2$.

Hyper-parameter	**Value**
Packing window (nominal)	131,072 (128K)
RoPE base θ	2×10^8 (ABF)
Training budget	2B tokens / domain
Global batch size	8 seq ($\approx$ 1M tokens)
Micro batch size	1
Peak learning rate	8×10^{-6}
Min learning rate	8×10^{-7}
LR schedule	cosine
Warmup ratio	10%
Optimizer	AdamW (β_1=0.9, β_2=0.95)
Weight decay	0.1
Gradient clipping	1.0
Precision	bfloat16
Attention backend	FlashAttention-3
Activation recomputation	full
Parallelism (TP/CP/PP/DP)	$2/4/1/4$
Hardware	4×8 H800 nodes (32 GPUs)
Training seeds	$1, 2, 42$

Table 10: Instruction-tuning hyper-parameters for the main Llama-3-8B-Base experiments, using UltraChat-200K with prompt-token masking (loss on responses only). The separate Qwen3-14B-Base RoPE-Filter scaling run uses TP/CP/PP/DP $= 4/4/1/2$ and its own tokenizer-specific packed cache.

Hyper-parameter	**Value**
Packing window (nominal)	131,072 (128K)
Peak learning rate	8×10^{-6}
Min learning rate	8×10^{-7}
LR schedule	cosine
Warmup ratio	10%
Global batch size	8 seq
Micro batch size	1
Epochs	1
Optimizer	AdamW
Precision	bfloat16
Activation recomputation	full
Training seeds	$1, 2, 42$
Hardware	4×8 H800 nodes (32 GPUs)

C Evaluation Benchmark Details

We evaluate two complementary axes. First, *long-context* ability on six benchmarks covering retrieval, understanding, reasoning, and long-form generation. Second, general *short-context* ability on five widely used benchmarks, which we treat as an ablation to confirm that long-context adaptation does not degrade general capability. Unless otherwise noted, scores are on a 0–100 scale and tasks within a benchmark are equally weighted. Table 11 summarizes the full suite, and the per-benchmark details follow, grouped into long- and short-context benchmarks.

Table 11: Overview of the eleven evaluation benchmarks used in this work: six long-context benchmarks and five short-context general-ability benchmarks. **#Smp.** denotes the number of evaluation samples; "Context" is the input-length regime.

Benchmark	Type	#Smp.	Context	Metric	Ability
Long-context benchmarks					
HELMET (Yen et al., 2025)	Real+synth.	9 tasks	32–128K	Mixed[†]	Retrieval, re-ranking, citation, recall, ICL
RULER (Hsieh et al., 2024)	Synthetic	3×500	4–128K	Recall	Tracing & aggregation
LongBench v2 (Bai et al., 2025)	Real	503	8K–2M w	Accuracy	Realistic reasoning
MRCR-v2 (OpenAI, 2025)	Synthetic	906	16–128K	Seq-match	Multi-needle retrieval
AA-LCR (Artificial Analysis, 2025)	Real	300	10–100K	LLM-judge	Long-document reasoning
LongProc (Ye et al., 2025)	Generation	1600	0.5–8K out	Rule-based	Procedural generation
Short-context benchmarks					
IFEval (Zhou et al., 2023)	Benchmark	541	Short	Rule-based	Instruction following
BBH (Suzgun et al., 2023)	Benchmark	6511	Short	Rule-based	Complex reasoning (CoT)
GSM8K (Cobbe et al., 2021)	In-house	1319	Short	Verifier	Grade-school math
MMLU-Pro (Wang et al., 2024)	Benchmark	12032	Short	Rule-based	Multi-subject knowledge
GPQA (Rein et al., 2024)	In-house	792	Short	Verifier	Expert-level science

[†] HELMET aggregates SubEM, NDCG@10, STR-EM/Rec@5, recall, and accuracy across its tasks (Table 12).

C.1 Long-context benchmarks

Table 12: Datasets and metrics of the HELMET subset used in this work, evaluated at 32K/64K/128K context lengths.

Category	Dataset	Metric	Description
Retrieval-augmented QA	PopQA	SubEM	Open-domain QA over Wikipedia
	JSON KV	SubEM	Key–value retrieval from JSON
Passage re-ranking	MS MARCO	NDCG@10	Re-rank passages for a query
Citation	ALCE ASQA	STR-EM	Cited answer generation
	ALCE QAMPARI	Rec@5	Multi-answer cited QA
Synthetic recall	RULER NIAH MK-2	Recall	Numeric needle, multi-key
	RULER NIAH MK-3	Recall	UUID needle, multi-key
	RULER NIAH MV	Recall	Multiple values per key
Many-shot ICL	CLINC150	Accuracy	150-way intent classification

HELMET. Following the HELMET protocol (Yen et al., 2025), we adopt nine representative tasks spanning retrieval-augmented QA (PopQA, JSON KV), passage re-ranking (MS MARCO), citation (ALCE ASQA/QAMPARI), synthetic recall (RULER NIAH MK-2/MK-3/MV), and many-shot in-context learning (CLINC150), evaluated at 32K/64K/128K (Table 12). Metrics are task-specific, including substring exact match (SubEM), NDCG@10, STR-EM/Rec@5 for citation, string-match recall, and accuracy.

RULER. RULER (Hsieh et al., 2024) is a fully synthetic, length-controllable stress test. Because its needle-in-a-haystack and QA families are already covered under the HELMET protocol, we additionally report Variable Tracking (VT), Common Words Extraction (CWE), and Frequent Words Extraction (FWE), evaluated at 4K–128K with 500 samples per task per length; scores use RULER's string-match recall.

LongBench v2. LongBench v2 (Bai et al., 2025) contains 503 multiple-choice questions with contexts from 8K to 2M words (mostly $<$128K), spanning single-document QA, multi-document QA, long in-context learning, long-dialogue understanding, code-repository understanding, and long structured-data understanding. We report accuracy, broken down by difficulty (easy/hard) and length (short/medium/long).

MRCR-v2. MRCR-v2 (OpenAI, 2025) is a multi-needle-in-a-haystack benchmark that requires identifying, disambiguating, and retrieving multiple highly similar needles in a long multi-turn dialogue. We use 906 samples of 16K–128K length and score by sequence-match ratio.

AA-LCR. AA-LCR (Artificial Analysis Long Context Reasoning) (Artificial Analysis, 2025) evaluates extraction, reasoning, and synthesis over long real-world documents (papers, financial and legal filings) of 10K–100K tokens, on which mid-2024 frontier models scored below 50%. We use its 100 questions, each evaluated over 3 repeats ($3 \times 100 = 300$ runs), scored with an LLM-as-a-Judge protocol (Qwen3-235B-A22B).

LongProc. LongProc (Ye et al., 2025) evaluates long-form procedural and structured generation, i.e., maintaining logical consistency and structural correctness in long outputs. It comprises 1600 questions with output-length tiers of 0.5K/2K/8K and rule-based scoring.

C.2 Short-context benchmarks

As a general-ability ablation, we verify that long-context adaptation does not degrade short-context ability by evaluating on five widely used English short-context benchmarks spanning instruction following, complex reasoning, mathematics, multi-subject knowledge, and expert-level science (summarized in Table 11). All are single-turn and scored automatically: IFEval measures adherence to programmatically verifiable format/length/keyword constraints; BBH collects 27 hard BIG-Bench tasks solved with chain-of-thought prompting; GSM8K tests multi-step grade-school math word problems; MMLU-Pro is a reasoning-heavy, 10-choice upgrade of MMLU; and GPQA poses PhD-level science questions that are hard to answer via search. GSM8K and GPQA are scored by an answer-verification system, and the others by rule-based metrics.