arXiv:2605.14589v3 [cs.CL] 5 Sep 2026

EndPrompt: Efficient Long-Context Extension via Terminal Anchoring

Han Tian[1,2*] **Luxuan Chen**[3,2*] **Xinran Chen**[2] **Rui Kong**[2] **Fang Wang**[2]
Jiamin Chen[2] **Jinman Zhao**[2] **Yuchen Li**[2†] **Jiashu Zhao**[2] **Shuaiqiang Wang**[2]
Haoyi Xiong[4] **Linghe Kong**[3] **Dawei Yin**[2†]

[1]Nankai University [2]Baidu Inc.
[3]Shanghai Jiao Tong University [4]Independent Researcher
tianhan@mail.nankai.edu.cn, yuchenli1230@gmail.com, yindawei@acm.org

Abstract

Extending the context window of large language models typically requires training on sequences at the target length, incurring quadratic memory and computational costs that make long-context adaptation expensive and difficult to reproduce. We propose **EndPrompt**, a method that achieves effective context extension using only short training sequences. The core insight is that exposing a model to long-range relative positional distances does not require constructing full-length inputs: we preserve the original short context as an intact first segment and append a brief terminal prompt as a second segment, assigning it positional indices near the target context length. This two-segment construction introduces both local and long-range relative distances within a short physical sequence while maintaining the semantic continuity of the training text–a property absent in chunk-based simulation approaches that split contiguous context. We provide a theoretical analysis grounded in Rotary Position Embedding and the Bernstein inequality, showing that position interpolation induces a rigorous smoothness constraint over the attention function, with shared Transformer parameters further suppressing unstable extrapolation to unobserved intermediate distances. Applied to LLaMA-family models extending the context window from 8K to 64K, EndPrompt achieves an average RULER score of 76.03 and the highest average on LongBench, surpassing LCEG (72.24), LongLoRA (72.95), and full-length fine-tuning (69.23) while requiring substantially less computation. These results demonstrate that long-context generalization can be induced from sparse positional supervision, challenging the prevailing assumption that dense long-sequence training is necessary for reliable context-window extension. The code is available at `https://github.com/clx1415926/EndPrompt`.

1 Introduction

Large language models are the foundation of modern natural language processing and a central interface for complex reasoning. However, their reliable maximum context length constrains utility. Applications such as long-document question answering [22], repository-level code understanding [21, 15], legal and scientific document analysis [5], personalized assistants, and extensive evidence retrieval require reasoning over inputs exceeding pretraining context windows. Here, longer contexts are effective only if models preserve local coherence and establish reliable interactions between distant tokens. Consequently, extending pretrained context windows is a central problem in language model adaptation.

Continuing to train models on long sequences at the target context length is a straightforward but costly solution. Collecting high-quality long-form corpora is difficult, and training increases memory

*Co-first authors with equal contributions.
†Corresponding author.

Preprint.

consumption and runtime due to the quadratic scaling of attention computation [29]. While efficient implementations like FlashAttention [12] and distributed systems mitigate this burden, full-length fine-tuning remains expensive. To reduce costs, recent methods explore position interpolation [8, 26, 6, 13], sparse attention [10], sliding-window attention [4, 31, 16], low-rank adaptation [19], and simulated long-context training [34, 2]. However, these approaches often introduce limitations, such as requiring substantial long-sequence training, altering the structure of attention, or chunking text in ways that damage semantic continuity. Thus, it remains unclear whether context extension requires dense supervision at the target length or if a sparse set of structured positional signals is sufficient.

This paper investigates whether models must observe full-length sequences to acquire long-context capabilities. We explore this question through the lens of positional generalization. In models utilizing Rotary Position Embedding [28], attention scores depend on both token content and relative positional distance. While existing methods assume reliable extrapolation requires exposure to dense relative distances during training, we demonstrate that effective long-context behavior emerges from sparse training signals. Specifically, models trained on short sequences can receive supervision for long-range positions if the examples preserve semantic coherence and provide stable anchors for distant positions. This finding reframes context adaptation as the design of informative positional supervision rather than merely increasing physical sequence length.

We propose **EndPrompt**, an efficient context-extension method utilizing positional index manipulation and an appended end prompt. We retain the original short context as the first segment and append a terminal prompt as the second. The first segment receives local positional indices, and the second receives indices near the target maximum context length. This configuration generates both local and long-range relative distances within a short physical sequence. Unlike chunk-based methods [34], our approach avoids splitting contiguous text, thereby preserving semantic integrity while exposing models to long-distance positional patterns. The end prompt functions as a stable terminal anchor. Experiments indicate robustness across various prompt formulations, suggesting efficacy derives from the terminal cue's structural position rather than memorizing specific tokens. This design exposes long-range positional interactions without compromising the quality of the short-context training signal.

We analyze how sparse supervision supports long-context generalization. Under Rotary Position Embedding [28], attention scores act as a sum of sinusoidal components over relative distances, featuring content-dependent amplitudes and phases. Position interpolation [8] reduces effective positional frequencies, constraining the attention score's variation rate and curvature across the distance dimension. This reduction induces a smoothness bias over unobserved intermediate distances. Furthermore, Transformers do not learn independent parameters for each distance; the same query and key projections support local and long-distance behavior. Consequently, sparse long-range supervision, combined with local supervision, constrains the shared parameter space and minimizes unstable behavior in unobserved regions. This theoretical perspective explains how short-sequence training produces stable long-context capabilities.

We evaluate **EndPrompt (ET)** on LLaMA-family models [30, 14], extending the context window from 8K to 64K. RULER [18] and LongBench [3] experiments show our method achieves competitive or superior performance compared to baselines such as LCEG [24], LongLoRA [9], and full-length fine-tuning. Specifically, our method achieves an average RULER score of 76.03, outperforming LCEG (72.24), LongLoRA (72.95), and full-length fine-tuning (69.23). On LongBench, our method secures the highest average score and demonstrates strong performance across tasks including question answering, summarization, few-shot learning, and code completion. Ablation studies validate the effects of end-prompt design, base model choice, extension length, and training-token quantity. These results confirm that reliable long-context adaptation emerges from structured sparse positional supervision without full-length training sequences.

The main contributions of this work are summarized as follows:

- We propose **EndPrompt (ET)**, an efficient context-extension method utilizing short training sequences, positional index manipulation, and an appended end prompt to simulate long-range positional supervision.
- We demonstrate that preserving the original context as an undivided segment maintains semantic continuity, while the end prompt provides a stable terminal anchor to create long-distance relationships without disrupting the signal for next-token prediction.

- We analyze the method through Rotary Position Embedding and position interpolation, explaining how smooth positional variation and shared parameters of the Transformer support generalization over unobserved intermediate distances.
- We demonstrate strong empirical performance on RULER and LongBench, where our approach outperforms representative baselines while avoiding full-length training sequences.

2 Preliminary

This section reviews the positional mechanisms of the proposed method: RoPE [28] and PI [8].

For a given attention head, let $\mathbf{q}_m, \mathbf{k}_n \in \mathbb{R}^D$ denote the query and key vectors at positions m and n. RoPE divides the dimensions into complex subspaces. In the j-th subspace, with assigned positional indices p_m and p_n, RoPE applies a position-dependent phase rotation to the content components $q_{m,j}$ and $k_{n,j}$. This yields $\hat{q}_{m,j} = q_{m,j}e^{ip_m\theta_j}$ and $\hat{k}_{n,j} = k_{n,j}e^{ip_n\theta_j}$, where θ_j is the fixed angular frequency. The unnormalized attention score contribution from this subspace depends on the assigned relative distance $d = p_m - p_n$. By expressing the content term $q_{m,j}\bar{k}_{n,j}$ in polar form with amplitude $a_j(\mathbf{x}_m, \mathbf{x}_n; \Theta)$ and phase offset $\phi_j(\mathbf{x}_m, \mathbf{x}_n; \Theta)$, the total RoPE attention score becomes a finite trigonometric polynomial:

$$S_{\mathrm{RoPE}}(d \mid \mathbf{x}_m, \mathbf{x}_n; \Theta) = \sum_{j=0}^{D/2-1} a_j(\mathbf{x}_m, \mathbf{x}_n; \Theta) \cos\left(d\theta_j + \phi_j(\mathbf{x}_m, \mathbf{x}_n; \Theta)\right). \tag{1}$$

Because the distance variable is exclusively embedded within the sinusoidal phase, modulating the positional indices enables attention over broader assigned relative distances without modifying the physical sequence length.

To adapt RoPE for extended contexts, PI rescales the positional indices by a target scale factor $s > 1$, mapping p to p/s. This operation effectively reduces the angular frequency to θ_j/s, modifying the overall attention score to:

$$S_{\mathrm{PI}}(d \mid \mathbf{x}_m, \mathbf{x}_n; \Theta) = \sum_{j=0}^{D/2-1} a_j(\mathbf{x}_m, \mathbf{x}_n; \Theta) \cos\left(\frac{d\theta_j}{s} + \phi_j(\mathbf{x}_m, \mathbf{x}_n; \Theta)\right). \tag{2}$$

Compared to Equation 1, this rescaling lowers the maximum rate of change along the distance dimension. Because S_{PI} is a finite trigonometric polynomial, the maximum effective frequency θ_0 strictly bounds the first-order variation and second-order curvature of the function:

$$\sup_d \left|\frac{\partial S_{\mathrm{PI}}}{\partial d}\right| \le \frac{\theta_0}{s} \sup_d |S_{\mathrm{PI}}(d)|, \quad \sup_d \left|\frac{\partial^2 S_{\mathrm{PI}}}{\partial d^2}\right| \le \left(\frac{\theta_0}{s}\right)^2 \sup_d |S_{\mathrm{PI}}(d)|. \tag{3}$$

Rather than guaranteeing perfect reconstruction for unseen distances, these bounds indicate that PI provides a smoothness bias by suppressing high-frequency positional variations. The proposed method utilizes this smoothness, combined with targeted long-distance supervision, to stabilize attention scores across distances unobserved during training.

3 Method

3.1 Overview

As illustrated in Figure 1, the proposed method aims to achieve efficient long-context adaptation without the high memory and computational costs associated with full-length sequence training. This objective is realized through two coupled components. First, positional index manipulation decouples the physical token order from the assigned positional indices, which creates sparse long-distance supervision while maintaining local attention. Second, an appended end prompt acts as a terminal anchor near the boundary of the target context window. This design preserves the semantic integrity of the original short context. Together, these components enable the model to acquire long-range positional capabilities from short training samples under the frameworks of RoPE and PI.

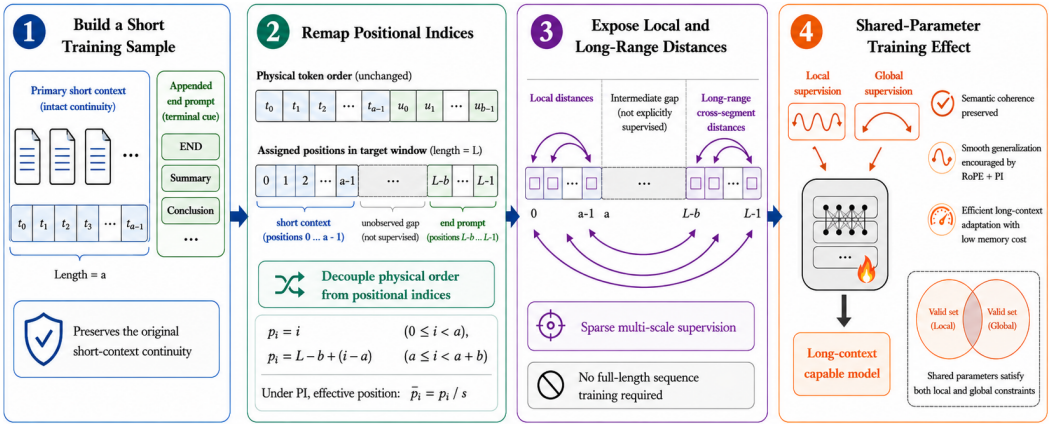


Figure 1: Overview of the proposed method.

3.2 Positional Index Manipulation

Let L denote the target context length. Given a short context sequence $\mathbf{x} = (x_0, x_1, \ldots, x_{a-1})$ of length a, an end prompt $\mathbf{e} = (e_0, e_1, \ldots, e_{b-1})$ of length b is appended to form the physical training sequence:

$$\mathbf{y} = (x_0, x_1, \ldots, x_{a-1}, e_0, e_1, \ldots, e_{b-1}), \quad |\mathbf{y}| = a + b. \tag{4}$$

While the physical length is $a + b$, the assigned positional indices span both ends of the target context window via the mapping:

$$p_\ell = \begin{cases} \ell, & 0 \le \ell < a, \\ L - b + (\ell - a), & a \le \ell < a + b. \end{cases} \tag{5}$$

Consequently, the short context and the end prompt are assigned to the intervals $[0, a-1]_{\mathbb{Z}}$ and $[L-b, L-1]_{\mathbb{Z}}$, respectively. With PI, the effective positional index becomes

$$\bar{p}_\ell = \frac{p_\ell}{s}, \tag{6}$$

for an interpolation scale factor $s > 1$. Thus, the attention score evaluates the assigned relative distance $p_\ell - p_r$ instead of the physical relative distance $\ell - r$:

$$S_{\ell r} = \sum_{j=0}^{D/2-1} a_{j,\ell r}(\Theta) \cos\left(\frac{(p_\ell - p_r)\theta_j}{s} + \phi_{j,\ell r}(\Theta) \right), \quad r \le \ell. \tag{7}$$

Equation 7 links the proposed position mapping with the PI score in Equation 2, allowing the attention mechanism to receive positional phases corresponding to the long-context range despite a short physical sequence.

Under causal attention, the observed set of assigned relative distances comprises local intervals from the original context and the end prompt, alongside a long-range interval between them:

$$D_{\text{obs}} = [0, a-1]_{\mathbb{Z}} \cup [0, b-1]_{\mathbb{Z}} \cup [L-a-b+1, L-1]_{\mathbb{Z}}. \tag{8}$$

Assuming $L - a - b \ge \max(a, b)$, the unobserved intermediate region is

$$D_{\text{gap}} = [\max(a, b), L-a-b]_{\mathbb{Z}}. \tag{9}$$

Rather than explicitly supervising all distances in D_{gap}, the model is trained on local and selected long distances. This mechanism relies on the smooth spectral structure of RoPE and PI to constrain behavior over the gap region, thereby providing sparse but multi-scale supervision for long-context adaptation.

3.3 End Prompt as the Terminal Segment

Splitting the original context to create long relative distances disrupts semantic continuity, as syntactic dependencies and local discourse relations rely on the original token order. Such splitting can remove essential local evidence for next-token prediction, degrading the quality of the supervision.

To circumvent this, the proposed method retains the intact original context and appends an end prompt as the terminal segment, assigned to the interval $[L-b, L-1]_{\mathbb{Z}}$ via Equation 5. This preserves local dependencies while establishing cross-segment distances approaching the target context length. The end prompt serves strictly as an explicit terminal cue rather than a semantic continuation.

Formally, the end prompt is sampled from a set of short terminal cues:

$$\mathbf{e} \sim \mathcal{E}, \tag{10}$$

where $\mathcal{E}$ denotes the cue set. A unique prompt string is unnecessary; the critical factor is structural placement near the end of the assigned context window. Provided the prompt offers a stable terminal cue without conflicting semantics, various formulations can induce the requisite long-distance interactions, thereby mitigating the risk of prompt memorization and enhancing robustness.

3.4 Training Objective

The proposed method integrates into standard autoregressive fine-tuning. Given the augmented sequence (Equation 4) and assigned positions (Equation 5), the training objective is

$$\mathcal{L}(\Theta) = -\sum_{\ell=0}^{a+b-2} w_\ell \log P_\Theta\left(y_{\ell+1} \mid y_{\leq \ell}; p_{\leq \ell}\right), \tag{11}$$

where $w_\ell \geq 0$ denotes an optional loss weight and $p_{\leq \ell}$ determines the attention phases. In practice, prompt-token losses are assigned a smaller but nonzero weight. This design reduces excessive reliance on prompt-token prediction while preserving the loss signal on terminal tokens, whose causal attention can attend to the original context over large assigned positional distances.

This optimization imposes both local and global constraints on shared Transformer parameters, preventing the model from learning independent parameters for each distance. Local constraints emerge from predictions within the original context and, for b, within the end prompt, whereas global constraints arise from the nonzero terminal-token losses, through which terminal-segment states attend to the original segment across large assigned distances.

Expressing these constraints through feasible parameter sets, the admissible region under purely local supervision is

$$\Theta_{\text{local}} = \left\{\Theta : \mathcal{L}_{\text{local}}(\Theta) \leq \varepsilon_{\text{local}}\right\}. \tag{12}$$

With terminal long-distance supervision, the region becomes

$$\Theta_{\text{valid}} = \left\{\Theta : \mathcal{L}_{\text{local}}(\Theta) \leq \varepsilon_{\text{local}}\right\} \cap \left\{\Theta : \mathcal{L}_{\text{global}}(\Theta) \leq \varepsilon_{\text{global}}\right\}. \tag{13}$$

This reduction in the feasible region eliminates parameter configurations that fail to generalize to long distances, acting as an implicit regularizer over the attention function (Equation 7).

3.5 Connection to Smooth Long-Context Adaptation

The effectiveness of the proposed method originates from the synergy between sparse distance exposure and the spectral properties of RoPE and PI. Because the manipulated positional indices alter only the phase term in Equation 7, the content-dependent amplitudes and phases remain governed by shared parameters. Consequently, the long-distance training signals act as a global constraint that directly regularizes the functions dictating local behavior. Furthermore, as indicated by Equation 3, PI reduces the effective angular frequency, which bounds the rate of positional variation and prevents unstable high-frequency oscillations within the unobserved gap region. In essence, the proposed method facilitates a constrained smooth extrapolation. The local and terminal supervisions anchor the short-range and long-range behaviors, PI suppresses excessive positional curvature, and the shared parameters unify these multi-scale constraints to achieve efficient long-context adaptation.

4 Experiments

4.1 Experimental Setup

The proposed method is evaluated on the architectures of LLaMA-2 7B [30] and LLaMA-3 8B [14]. The default configuration utilizes a corpus of one billion tokens to extend the context window

from 8K to 64K. LongBench [3] and RULER [18] are utilized to evaluate the capabilities of the models in processing extended contexts. Furthermore, standard benchmarks, including GSM8K [11], HumanEval [7], MMLU [17], and HellaSwag [32], are employed to assess the capabilities for short-text understanding. A comprehensive description of the evaluation tasks and the specific datasets is provided in Appendix A.2.

Baselines We compare the proposed approach against four strong baselines: Positional Skip-Embedding [34], LCEG [24], LongLoRA [9], and full-length fine-tuning. Positional Skip-Embedding extends the context window by chunking inputs and manipulating position indices within a fixed window. LCEG provides a standardized protocol for evaluating the generalization of long contexts. LongLoRA accelerates the extension process using shifted sparse attention to minimize computational costs while retaining the original architectures. Finally, full-length fine-tuning trains the models directly on the target context length, serving as a resource-intensive standard for comparison.

Table 1: Results on the RULER synthetic benchmark (4K-64K).

Task	Method			
	ET	LCEG	LongLoRA	Full FT
Niah_S1	100.00	100.00	100.00	97.56
Niah_S2	91.28	99.28	99.44	98.12
Niah_S3	92.92	79.68	86.20	72.72
Niah_M1	90.20	96.12	97.36	94.52
Niah_M2	85.48	76.48	77.96	90.24
Niah_M3	62.92	45.72	51.92	56.20
Niah_MV	81.67	77.81	81.56	62.34
Niah_MQ	82.06	83.88	83.79	62.56
Vt	82.00	68.18	65.70	68.56
Cwe	42.82	53.14	50.97	38.32
Fwe	83.53	63.01	58.17	58.10
Qa_1	51.88	53.04	52.40	55.68
Qa_2	41.60	42.84	42.92	45.04
Avg.	**76.03**	72.24	72.95	69.23

Table 2: Results on various downstream tasks.

Method	Single-Doc QA	Multi-Doc QA	Summarization	Few-Shot Learning	Synthetic Task	Code Completion	Avg. Score
ET	**32.03**	**30.81**	**26.04**	**68.04**	4.54	**66.48**	**38.30**
LCEG	27.86	28.51	23.88	61.81	5.31	46.86	36.61
LongLoRA	26.86	27.51	22.88	60.81	4.31	45.86	36.84
Full FT	28.86	29.51	24.88	62.81	**6.31**	47.86	35.63

4.2 Main Results on Long-Context Benchmarks

Table 1 and Table 2 present the performance of the models on RULER and LongBench. In the main comparison, models are trained on a one-billion-token corpus to extend the context window from 8K to 64K. The results demonstrate the ability of the proposed method to achieve superior performance in long-document understanding and retrieval compared to full-length fine-tuning and parameter-efficient baselines. Furthermore, we evaluate the training efficacy of the proposed method in terms of memory footprint and time consumption. The results indicate that our method effectively overcomes the traditional space-time trade-off, achieving significant reductions in memory utilization while simultaneously accelerating training speed compared to baseline methodologies. Detailed results and comprehensive analysis can be found in Appendix B.2.

Superior Overall Performance across Benchmarks ET consistently achieves the highest average performance across both frameworks. On RULER, ET reaches an average score of 76.03, outperforming LongLoRA (72.95) and LCEG (72.24). On LongBench, the standard ET secures an average score

of 38.30, exceeding full-length fine-tuning (35.63). This consistent gap highlights the effectiveness of ET across varied context lengths and tasks.

Resilience Divergence in Complex Information Retrieval While the single-needle retrieval tasks (e.g., Niah_S1) saturate at perfect scores of 100.00 for ET, LCEG, and LongLoRA, the performance diverges significantly as the complexity increases. ET demonstrates substantial robustness in demanding settings, outperforming LongLoRA on Vt (82.00 vs. 65.70) and Fwe (83.53 vs. 58.17). In multi-needle scenarios (Niah_MV and Niah_MQ), ET maintains strong scores of 81.67 and 82.06, respectively. These results indicate that the proposed method extends the context window without degrading the capacity to extract deeply embedded information.

Profound Advantages in Practical Downstream Applications ET exhibits significant advantages in practical downstream applications. It achieves the highest scores across all realistic text processing domains, notably in Code Completion (66.48), significantly exceeding LCEG (46.86) and LongLoRA (45.86). A similar margin is observed in Few-Shot Learning, where ET scores 68.04 compared to a baseline average of approximately 61. By leading in Single-Doc QA, Multi-Doc QA, and Summarization, ET demonstrates that the architectural enhancements effectively translate to real-world semantic reasoning.

Table 3: Ablation study on the RULER synthetic benchmark across different configurations.

Method	Task								Avg.
	Niah_Single	Niah_Multikey	Niah_Multivalue	Niah_Multiquery	Vt	Cwe	Fwe	Qa	
Main									
ET	94.73	79.53	81.67	82.06	82.00	42.82	83.53	46.74	76.03
Different Base Models									
Mistral-7B-v0.3	98.75	70.92	62.07	87.36	61.46	19.09	59.76	45.16	68.39
LLaMA-2-7B	66.08	38.91	48.78	55.92	54.80	21.66	36.80	31.34	45.82
Different Extension Lengths									
ET-32k	98.23	80.00	86.65	93.45	78.03	44.85	85.58	47.73	78.36
ET-96k	97.21	72.12	86.21	91.41	86.86	41.83	83.64	45.74	76.11
ET-128k	99.20	73.31	53.21	76.74	84.24	39.16	83.11	46.36	72.82
Different Token Quantities									
ET-2.0B	95.65	79.05	85.60	85.33	70.49	42.63	79.08	47.82	75.61
ET-0.5B	94.71	74.35	89.53	88.71	78.58	38.24	83.11	43.04	74.72

Table 4: Ablation study on LongBench across different configurations.

Method	Task						Avg.
	Single-Doc QA	Multi-Doc QA	Summarization	Few-Shot Learning	Synthetic Task	Code Completion	
Main							
ET	32.03	30.81	26.04	68.04	4.54	66.48	38.30
Different Base Models							
Mistral-7B-v0.3	34.90	32.76	23.88	65.79	3.25	59.05	37.29
LLaMA-2-7B	25.86	26.51	21.88	59.81	3.31	44.86	31.16
Different Extension Lengths							
ET-128k	30.18	28.41	22.68	67.21	2.27	60.41	35.68
ET-96k	31.51	29.80	24.90	67.89	5.43	58.49	36.88
ET-32k	30.30	26.23	26.19	68.42	4.50	60.39	36.45
Different Token Quantities							
ET-2.0B	29.03	29.61	25.37	66.29	14.00	63.44	37.86
ET-0.5B	28.67	27.86	25.80	67.83	12.29	65.27	37.85

4.3 Ablation Studies

In this subsection, we conduct extensive ablation studies to evaluate the robustness and scalability of the proposed method across diverse configurations. First, we verify the broad applicability of the proposed approach across different base models, including the architecture of Mistral [20]. Second, we explore the retention of performance when the context window is expanded to extreme lengths of up to 128K tokens. Third, we investigate the impact of the volume of training data on the consistent scaling of performance. Finally, we examine the sensitivity of the framework to specific variations

of the end prompt. The experimental results systematically deconstruct the impact of each factor, demonstrating the stability of the method under varying conditions.

Broad Applicability across Diverse Model Families The proposed method demonstrates broad generalizability across distinct foundation architectures. When integrated into Mistral-7B-v0.3 and LLaMA-2 7B, the framework achieves consistent contextual scaling. The architecture of Mistral reaches average scores of 68.39 on the benchmark of RULER and 37.29 on LongBench, whereas the architecture of LLaMA-2 scores 45.82 and 31.16, respectively. This variance is attributable to the pre-training quality of the underlying models rather than any limitation of the methodology. Thus, the extension technique functions as an effective, model-agnostic approach for long-context processing.

Robust Performance Retention across Extended Context Windows The framework extends the context window to various lengths while preserving high overall performance. On the benchmark of RULER, the model achieves average scores of 78.36 at the length of 32K, 76.11 at 96K, and 72.82 at the extreme context of 128K. This stability is mirrored in LongBench, with scores ranging from 36.45 at 32K to 35.68 at 128K. The minor degradation of performance at extreme lengths is expected given the substantial expansion of the context window, confirming the reliable scalability without catastrophic forgetting.

Consistent Performance Scaling with Increased Training Tokens Scaling the training corpus from 0.5 billion to 2.0 billion tokens yields consistent performance enhancements. The average score increases from 74.72 to 75.61 on the benchmark of RULER, and from 37.85 to 37.86 on LongBench. Specifically, the configuration of 2.0 billion tokens achieves a score of 29.61 in Multi-Doc QA, compared to 27.86 for the setup of 0.5 billion tokens. These results indicate that the method effectively utilizes increased volumes of data to fortify the long-context capabilities.

Consistent Robustness across End Prompt Variations The proposed framework maintains stable performance regardless of the configuration of the end prompt. On the benchmark of LongBench, the average score exhibits a negligible decline from 38.30 (EP_1) to 37.95 (EP_3). Similarly, on RULER, the performance ranges from 76.45 (EP_2) to 74.63 (EP_3). These minimal fluctuations demonstrate that the extension mechanism is robust and insensitive to specific instructional wording, effectively leveraging generalized termination signals to minimize the need for extensive prompt engineering. A comprehensive description of the specific end prompts is provided in Appendix B.1.

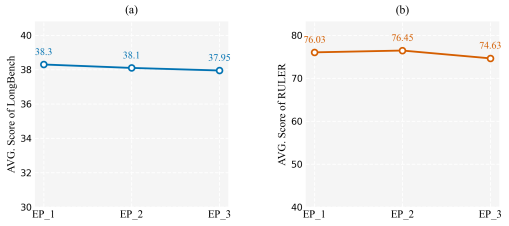


Figure 2: Ablation study on different end prompts.

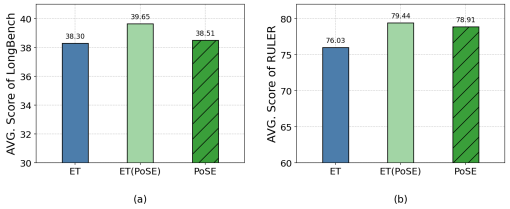


Figure 3: Performance comparison among the standard ET, Positional Skip-Embedding, and ET(PoSE) across the benchmarks of LongBench and RULER.

4.4 Structural Analysis

To evaluate compatibility, the proposed approach is compared against Positional Skip-Embedding, a method that extends the context window by partitioning inputs into chunks and manipulating position indices to simulate extended lengths. We integrate this training methodology with the proposed framework, denoting the hybrid configuration as ET(PoSE). The performances of the standard ET, Positional Skip-Embedding, and the integrated ET(PoSE) are systematically evaluated on LongBench and RULER. As shown in Figure 3, the proposed method demonstrates strong compatibility with Positional Skip-Embedding. On LongBench, the standard ET achieves an average score of 38.30, which is comparable to the 38.51 obtained by Positional Skip-Embedding. On RULER, ET scores 76.03, competitive with the 78.91 achieved by the baseline. Crucially, the hybrid ET(PoSE) configuration yields the highest performance, reaching peak scores of 39.65 on LongBench

and 79.44 on RULER. These findings confirm that leveraging the chunking strategy of Positional Skip-Embedding within the ET framework significantly enhances the overall performance.

4.5 Evaluation and Recovery of Short-Text Capabilities

This subsection investigates the impact of context extension on the short-text capabilities using the LLaMA-3 8B architecture. The models are evaluated on standard benchmarks, including GSM8K [11], HumanEval [7], MMLU [17], and HellaSwag [32]. To mitigate performance degradation, a supervised fine-tuning phase is applied. Table 5 presents the recovered capabilities. The proposed approach, denoted as sft_ET(PoSE), achieves the highest average score of 53.56, surpassing full fine-tuning (50.98), LongLoRA (48.64), and LCEG (49.74). Without the integration of Positional Skip-Embedding, the standard ET method still secures a strong average of 52.41. While standalone Positional Skip-Embedding achieves 52.32, the performance discrepancy highlights the effects of the differing training mechanisms. Standard Positional Skip-Embedding partitions the actual training data, risking the disruption of semantic continuity. Conversely, the proposed method partitions only the end prompts, preserving the structural and semantic integrity of the text and ensuring superior performance on short-text tasks.

Table 5: Results on short-text evaluation tasks after supervised fine-tuning.

Method	**Variant**	**Evaluation Task**				**Avg.**
		MMLU	GSM8K	HS	HE	
SFT	sft_ET(PoSE)	**56.87**	**46.10**	**78.34**	**32.93**	**53.56**
	sft_ET	<u>56.67</u>	<u>44.66</u>	77.51	30.83	<u>52.41</u>
	sft_LongLoRA	55.07	38.82	77.49	23.17	48.64
	sft_LCEG	54.76	42.23	77.56	24.39	49.74
	sft_PoSE	56.62	43.59	<u>77.95</u>	<u>31.10</u>	52.32
	sft_Full FT	55.40	42.15	77.11	29.27	50.98

5 Conclusion

We presented EndPrompt, an efficient approach for extending the context window of large language models without training on full-length sequences. The method preserves the original short context as an intact segment and appends a brief terminal prompt whose positional indices are placed near the target context boundary. This construction exposes the model to both local and long-range relative distances while avoiding the semantic disruption introduced by chunk-based simulation. Our analysis explains this behavior through RoPE and position interpolation: RoPE represents relative distance through a shared spectral basis, PI suppresses high-frequency positional variation, and shared Transformer parameters couple local and terminal long-distance supervision within the same attention function. Empirically, EndPrompt extends LLaMA-family models from 8K to 64K context and achieves strong results on RULER and LongBench, including a 76.03 average RULER score that exceeds the reported LCEG, LongLoRA, and full-length fine-tuning baselines. These findings suggest that reliable long-context adaptation can be induced from structured sparse positional supervision rather than dense full-length training.

The current formulation relies on an explicit terminal segment. Our future work would study how terminal anchoring can be combined with streaming, hierarchical, or multimodal long-context architectures. More broadly, EndPrompt points to a practical direction for context-window extension – designing informative positional supervision while keeping the physical training sequence short.

References

[1] Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, 2024.

[2] Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1376–1395, 2024.

[3] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3119–3137, 2024.

[4] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[5] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, 2022.

[6] Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, and Lidong Bing. Clex: Continuous length extrapolation for large language models. In *International Conference on Learning Representations*, volume 2024, pages 14352–14369, 2024.

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[8] 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.

[9] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. In *International Conference on Learning Representations*, volume 2024, pages 8220–8238, 2024.

[10] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[11] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[12] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

[13] 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. *arXiv preprint arXiv:2402.13753*, 2024.

[14] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[15] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yifan Wu, YK Li, et al. Deepseek-coder: when the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

[16] Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language 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)*, pages 3991–4008, 2024.

[17] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[18] 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.

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[20] Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10(3), 2023.

[21] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pages 54107–54157, 2024.

[22] Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

[23] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ringattention with blockwise transformers for near-infinite context. In *International Conference on Learning Representations*, volume 2024, pages 3992–4008, 2024.

[24] Yi Lu, Jing Nathan Yan, Songlin Yang, Justin T Chiu, Siyu Ren, Fei Yuan, Wenting Zhao, Zhiyong Wu, and Alexander M Rush. A controlled study on long context extension and generalization in llms. *arXiv preprint arXiv:2409.12181*, 2024.

[25] Bowen Peng and Jeffrey Quesnelle. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation, 2023.

[26] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pages 31932–31951, 2024.

[27] Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, 2023.

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

[29] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

[30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[31] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pages 21875–21895, 2024.

[32] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800, 2019.

[33] Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. Soaring from 4k to 400k: Extending llm's context with activation beacon. *arXiv preprint arXiv:2401.03462*, 2(3):5, 2024.

[34] 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 *International Conference on Learning Representations*, volume 2024, pages 18940–18954, 2024.

A Technical Appendices

A.1 Experiment details

We adopt Meta-Llama-3-8B as our base model and employ Position Interpolation (PI) to extend its context window. During data preprocessing, the maximum sequence length of the training corpus is 8K. We perform full-parameter fine-tuning for one epoch on a single computing node equipped with 8 NVIDIA A800 (80GB) GPUs. To optimize memory consumption and training efficiency, all training procedures utilize BF16 mixed precision, integrated with DeepSpeed ZeRO Stage-3, FlashAttention, and gradient checkpointing. For the optimization hyperparameters, we set the peak learning rate to 2×10^{-5} and utilize a constant learning rate scheduler with a 20-step linear warmup.

A.2 Evaluation Tasks

To thoroughly assess the performance of the models, the experiments incorporate a diverse array of datasets across multiple evaluation frameworks. LongBench [3] provides a comprehensive multi-task evaluation encompassing datasets for single-document and multi-document question answering, document summarization, few-shot learning, synthetic tasks, and code completion. Additionally, RULER [18] is utilized to conduct a fine-grained analysis of the effective context length. This benchmark evaluates the information retrieval capabilities of the models through specific dataset configurations, including multi-needle-in-a-haystack tasks, variable tracking, common word extraction, frequent word extraction, and synthetic question answering. These datasets systematically test the extraction of target information embedded within extensive background texts across varying sizes of the context window. Finally, the evaluation of short-text capabilities employs GSM8K [11] for mathematical reasoning, HumanEval [7] for code generation, MMLU [17] for massive multi-task language understanding, and HellaSwag [32] for commonsense reasoning.

B Supplementary Experiment

B.1 Detailed End Prompts

This section details the specific end prompts evaluated in the robustness analysis. The first configuration, denoted as EP_1, utilizes the explicit phrase "This is the end of text, please pay attention here". The second configuration, EP_2, employs the native special termination token of the LLaMA-3 architecture, represented as `<|eot_id|>`. Finally, the third configuration, EP_3, consists of the minimal string "End.".

B.2 Training Efficacy

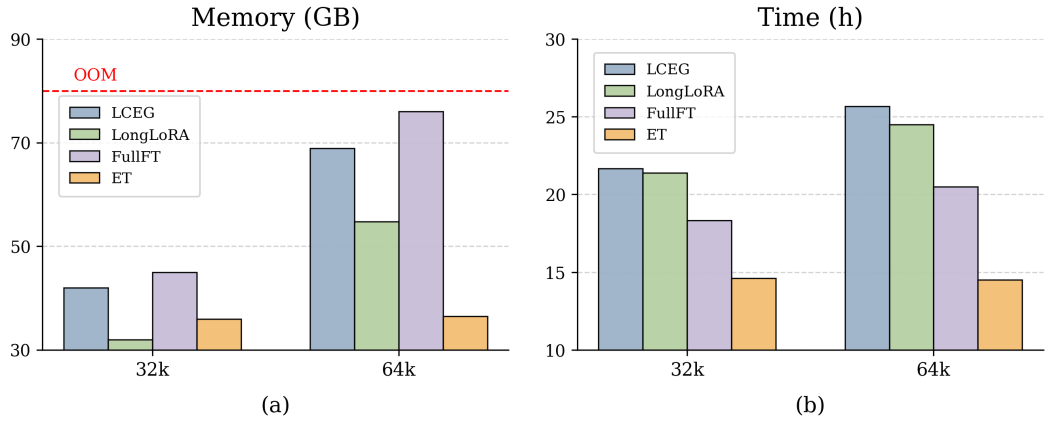


Figure 4: Comparison of memory footprint and time consumption across different methods.

This subsection analyzes the training efficiency by evaluating the memory footprint and the temporal cost. As illustrated in Figure 4, the ET method overcomes the traditional space-time trade-off,

achieving strict dominance over the baseline methodologies at the context lengths of 32K and 64K. At 64K, ET requires only 36.52 GB of memory, representing a 52% reduction compared to the 76.00 GB required by full-length fine-tuning. Concurrently, ET achieves acceleration ratios of 1.41×, 1.69×, and 1.77× over full-length fine-tuning, LongLoRA, and LCEG, respectively. This dual advantage of reduced memory utilization and increased computational speed validates the efficiency of the method in scaling context windows without prohibitive overhead.

B.3 Complete Table1

Table 6: Results on the RULER synthetic benchmark (4K).

Task	**Method**			
	ET	**LCEG**	**LongLoRA**	**Full FT**
Niah_S1	100.00	100.00	100.00	96.40
Niah_S2	100.00	98.40	98.40	99.40
Niah_S3	99.40	71.00	83.40	71.20
Niah_M1	97.00	97.40	98.80	98.00
Niah_M2	99.20	93.60	92.80	99.40
Niah_M3	96.80	72.80	80.80	83.80
Niah_MV	97.20	88.90	92.40	75.40
Niah_MQ	99.15	97.45	97.20	84.50
Vt	99.48	95.96	93.76	93.20
Cwe	91.82	88.60	83.36	82.50
Fwe	90.07	62.40	61.13	76.90
Qa_1	64.60	57.80	57.20	66.40
Qa_2	47.20	47.00	46.00	51.80
Avg.	**90.92**	82.41	83.48	82.99

Table 7: Results on the RULER synthetic benchmark (8K).

Task	**Method**			
	ET	**LCEG**	**LongLoRA**	**Full FT**
Niah_S1	100.00	100.00	100.00	98.00
Niah_S2	100.00	100.00	100.00	99.00
Niah_S3	98.80	89.20	93.40	72.60
Niah_M1	98.80	96.20	97.40	97.60
Niah_M2	99.20	75.40	83.20	98.60
Niah_M3	92.00	59.20	63.20	78.00
Niah_MV	96.25	93.00	94.60	87.80
Niah_MQ	96.85	96.50	97.30	87.50
Vt	99.20	85.20	78.96	85.00
Cwe	65.50	68.66	65.00	55.80
Fwe	80.67	67.13	66.13	57.80
Qa_1	59.40	53.80	53.40	57.80
Qa_2	45.20	46.40	46.00	50.60
Avg.	**87.07**	79.28	79.89	78.93

Table 8: Results on the RULER synthetic benchmark (16K).

Task	Method			
	ET	**LCEG**	**LongLoRA**	**Full FT**
Niah_S1	100.00	100.00	100.00	98.00
Niah_S2	100.00	100.00	100.00	96.40
Niah_S3	97.20	91.60	93.80	79.20
Niah_M1	95.40	95.20	96.60	95.60
Niah_M2	95.00	63.80	71.00	95.80
Niah_M3	67.00	34.60	48.00	62.60
Niah_MV	95.30	89.45	92.75	54.60
Niah_MQ	92.40	93.85	94.10	45.80
Vt	94.52	79.88	72.96	74.20
Cwe	49.76	61.94	61.22	33.30
Fwe	90.00	71.07	65.53	67.20
Qa_1	61.00	53.60	53.20	58.60
Qa_2	45.00	41.80	42.60	47.80
Avg.	**83.28**	75.14	76.29	69.93

Table 9: Results on the RULER synthetic benchmark (32K).

Task	Method			
	ET	**LCEG**	**LongLoRA**	**Full FT**
Niah_S1	100.00	100.00	100.00	98.00
Niah_S2	100.00	99.80	100.00	99.20
Niah_S3	98.00	84.60	87.80	78.40
Niah_M1	92.20	98.60	98.60	97.40
Niah_M2	77.80	79.80	77.00	88.80
Niah_M3	53.00	35.60	40.00	38.60
Niah_MV	93.75	74.85	79.60	65.30
Niah_MQ	91.10	83.65	83.45	64.90
Vt	93.36	66.96	56.96	55.90
Cwe	6.00	41.98	39.32	14.10
Fwe	88.13	61.27	55.60	44.70
Qa_1	44.40	52.60	52.20	52.80
Qa_2	40.20	45.20	43.40	42.20
Avg.	**75.23**	71.15	70.30	64.64

C Related Work

Position-based Context Extension Large language models predominantly employ Rotary Position Embedding (RoPE) [28] to encode positional information. To extend the context length beyond the pretraining limit, Position Interpolation (PI) [8] rescales the positional indices to fit within the original domain. Further improvements, such as NTK-aware scaling [25] and YaRN [26], modify the rotary frequencies based on the neural tangent kernel theory to preserve high-frequency local information while interpolating low-frequency long-range information. While effective, these methods still typically require fine-tuning on long text sequences to achieve optimal performance and alignment.

Efficient Long-Context Adaptation Training on full-length sequences incurs quadratic computational complexity. To mitigate this, LongLoRA [9] introduces shifted sparse attention, enabling efficient fine-tuning without significantly modifying the core architecture. Alternatively, simulated long-context training methods like PoSE [34] manipulate positional indices across disjoint chunks of text to simulate long distances. Other orthogonal approaches include RingAttention [23] for distributed sequence processing across multiple devices, and Activation Beacons [33] that compress context into condensed representations. In contrast to chunking methods that may break semantic continuity, our approach maintains the intact short context and utilizes an explicit end prompt to anchor long-range positional indices efficiently.

Long-Context Evaluation Evaluating extended context capabilities requires rigorous and standardized benchmarks. LongBench [3] and L-Eval [1] offer diverse multi-task suites encompassing summarization, question answering, and code analysis. ZeroSCROLLS [27] focuses specifically on zero-shot long text understanding for summarization tasks. More recently, RULER [18] provides a configurable synthetic evaluation framework to test the effective context limits of LLMs by escalating the complexity of needle-in-a-haystack tasks with multi-key and multi-value retrieval scenarios.