Longer Context, Deeper Thinking: Uncovering the Role of Long-Context Ability in Reasoning

Wang Yang[1], **Zirui Liu**[2], **Hongye Jin**[3], **Qingyu Yin**
Vipin Chaudhary[1], **Xiaotian Han**[1]
[1]Case Western Reserve University [2] University of Minnesota - Twin Cities
[3]Texas A&M University
{wxy320,vipin,xhan}@case.edu, zrliu@umn.edu, jhy0410@tamu.edu

Abstract

Recent language models exhibit strong reasoning capabilities, yet the influence of long-context capacity on reasoning remains underexplored. In this work, we hypothesize that current limitations in reasoning stem, in part, from insufficient long-context capacity, motivated by empirical observations such as *i)* higher context window length often leads to stronger reasoning performance, and *ii)* failed reasoning cases resemble failed long-context cases. To test this hypothesis, we examine whether enhancing a model's long-context ability before Supervised Fine-Tuning (SFT) leads to improved reasoning performance. Specifically, we compared models with identical architectures and fine-tuning data but varying levels of long-context capacity. Our results reveal a consistent trend: models with stronger long-context capacity achieve significantly higher accuracy on reasoning benchmarks after SFT. Notably, these gains persist even on tasks with short input lengths, indicating that long-context training offers generalizable benefits for reasoning performance. These findings suggest that long-context modeling is not just essential for processing lengthy inputs, but also serves as a critical foundation for reasoning. We advocate for treating long-context capacity as a first-class objective in the design of future language models. Our code is anonymously available at https://github.com/uservan/LCTMerge.

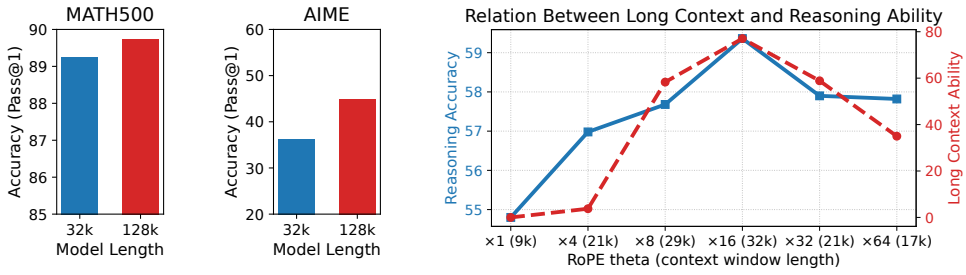


Figure 1: Impact of long-context capacity on mathematical reasoning. **Left:** Accuracy (Pass@1) on MATH500 and AIME datasets for public models with 32k and 128k context lengths, showing consistent improvements in reasoning performance with longer context windows. The 32k and 128k LLMs refer to three different public models, as shown in Table 1. **Right:** Reasoning accuracy versus RoPE theta values, highlighting a strong correlation between long-context capacity and reasoning performance. Increasing the RoPE theta value typically extends the effective context window length.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Large language models (LLMs) have recently demonstrated impressive reasoning capabilities across a wide range of benchmarks [1–3]. Despite this progress, the underlying factors that influence such reasoning abilities remain only partially understood. One particularly underexplored dimension is long-context ability—the model's capacity to utilize a longer reasoning path during inference—which could affect the reasoning performances and the reasoning model fine-tuning.

While prior work has primarily focused on training paradigms and dataset quality to enhance reasoning [4–6], we hypothesize that a model's reasoning ability is also fundamentally constrained by its long-context capacity. This hypothesis is grounded in three empirical observations. *First*, models with extended context lengths (e.g., 128k vs. 32k) consistently achieve higher accuracy on reasoning benchmarks such as MATH500 and AIME, suggesting a direct performance benefit from stronger long-context modeling (Table 1, Figure 1). *Second*, case studies reveal that failed generations often involve extremely long outputs, sometimes truncated at generation limits, and exhibit issues like repetition or incorrect cross-referencing—failure patterns strongly linked to inadequate long-context capability (Figure 2, Figure 3). *Third*, modern reasoning datasets now include many samples exceeding 8K or even 10K tokens—substantially longer than early CoT data—requiring models to learn from long, variable-length reasoning sequences (Figure 4). Together, these findings underscore long-context capacity as a critical factor for reasoning ability, and motivate the central question: *Does improving a model's long-context ability during pretraining enhance downstream reasoning?*

To rigorously investigate this, we conduct a controlled study comparing language models with identical architectures and fine-tuning data, but varying degrees of long-context pretraining. Our experimental results reveal a consistent and compelling trend: models with stronger long-context capabilities consistently outperform their counterparts on reasoning tasks after SFT. Notably, these improvements extend to reasoning problems with short input lengths, suggesting that long-context training imparts generalizable cognitive benefits that go beyond simply processing long sequences. As shown in Figure 1, `LLaMA3-8B-Instruct` exhibits varying reasoning performance after training when equipped with different levels of long-context capability. Notably, reasoning ability tends to increase or decrease in accordance with the strength of the model's long-context capacity, suggesting a direct correlation between the model's reasoning ability and long-context capacity.

Based on our experimental results, we thus propose a *Recipe for Reasoning Fine-Tuning*, which advocates for *appropriately enhancing a model's long-context capacity prior to reasoning SFT*—for instance, by extending its context length to 128K tokens. Applying this recipe to `Qwen2.5-Math-7B-Instruct`, we observe substantial improvements: performance on `MATH500` increases from an average of 85.04 to 88.70, and on `AIME`, from 15.00 to 28.00.

2 Motivation: behavioral evidence suggests a connection between long context and reasoning

In this section, we present a set of empirical observations that suggest a strong behavioral connection between a model's long-context capability and its reasoning performance. Through controlled comparisons, token length analyses, and failure case studies, we observe that LLMs with stronger long-context abilities not only perform better on reasoning benchmarks but also handle diverse and extended reasoning sequences more reliably. These findings collectively highlight long-context modeling as a key factor and component in enabling strong reasoning ability.

2.1 Higher context window length often leads to stronger reasoning performance.

Recent advances in long-context modeling have enabled language models (LLMs) to process substantially longer sequences. However, it remains unclear whether such long-context capacities yield tangible benefits for reasoning tasks. In this section, we empirically examine the relationship between long-context ability and reasoning performance. We collect a set of well-known open-source reasoning models fine-tuned from `Qwen/Qwen2.5-7B-Instruct`. These models are categorized into two groups based on their long-context capacity: 32k and 128k tokens. We then evaluate and compare their reasoning performance on two math reasoning benchmarks: `MATH500` and `AIME`.

The detailed results are reported in Table 1. Figure 1 presents that models with longer context lengths (128k vs. 32k) consistently achieve higher accuracy on mathematical reasoning benchmarks such as MATH500 and AIME. This suggests that the ability to encode and maintain longer contextual dependencies can directly translate into better reasoning capabilities. These results collectively highlight the importance of effective long-context training—not only for tasks involving long inputs, but also for general reasoning even when test-time inputs are relatively short.

Table 1: Performance comparison on MATH500 and AIME benchmarks for some popular open-source reasoning models with different long-context abilities at 32k and 128k context lengths. Reasoning models with enhanced long-context capacity (128k) generally exhibit improved performance, particularly on the AIME benchmark. Averages are reported in the bottom row.

Long Context Ability at 32k	MATH500	AIME	Long Context Ability at 128k	MATH500	AIME
OpenR1-Qwen-7B	90.36	43.11	DeepSeek-R1-Distill-Qwen-7B	91.68	45.56
OpenThinker-7B	86.80	25.78	OpenMath-Nemotron-7B	94.00	74.67
OpenThinker2-7B	90.60	40.22	DeepMath-Zero-7B	83.48	14.67
OpenThinker3-7B	93.72	64.00	AceReason-Nemotron-7B	93.84	62.89
Avg (32k)	90.37	43.28	Avg (128k)	90.75	49.45

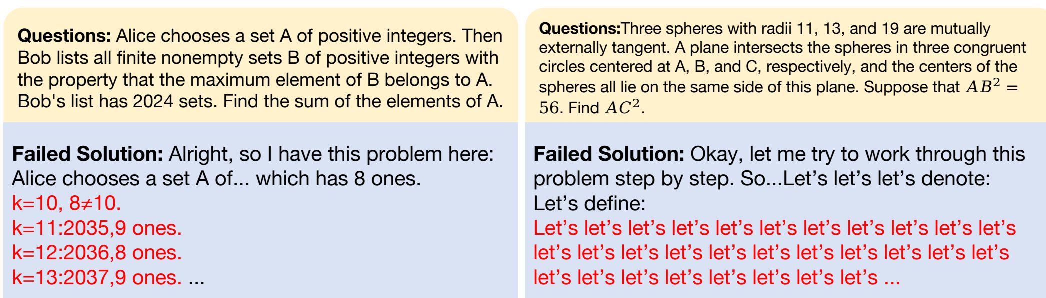


Figure 2: **Case Study: Repetition Failure.** Two failure cases where the model produces clearly repetitive sentences in its answers. Such repetition is a common symptom of insufficient long-context capability, leading to strange responses and degraded reasoning quality in extended sequences.

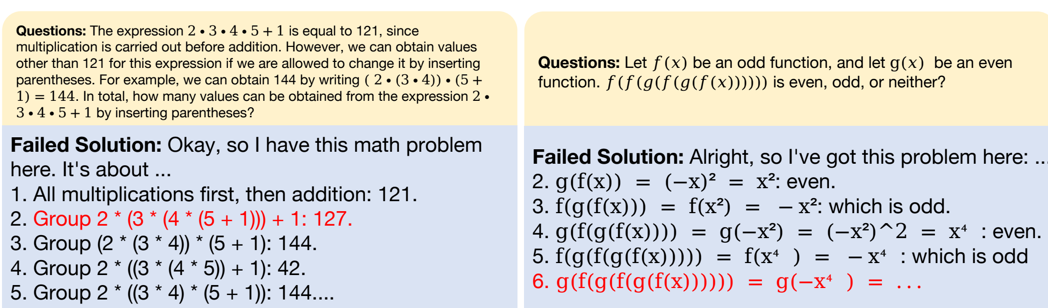


Figure 3: **Case Study: Contextual Reference Failures.** Two failure cases where the model makes incorrect references to expressions introduced earlier in the problem. These errors occur in the later stages of the response and reflect a typical symptom of insufficient long-context capability.

2.2 Case study: failed reasoning cases resemble failed long-context cases.

We analyze failed cases of reasoning models and long-context inputs, revealing a strong connection between reasoning ability and long-context capacity. Specifically, we examine the average input lengths of correct and incorrect predictions made by three reasoning models—`DeepSeek-R1-Distill-Qwen-1.5B`, `7B`, and `14B`—on two math benchmarks: `MATH500` and `AIME`. As shown in Figure 4, incorrect predictions are typically associated with inputs exceeding 10k tokens, suggesting that these failures may stem from insufficient long-context handling.

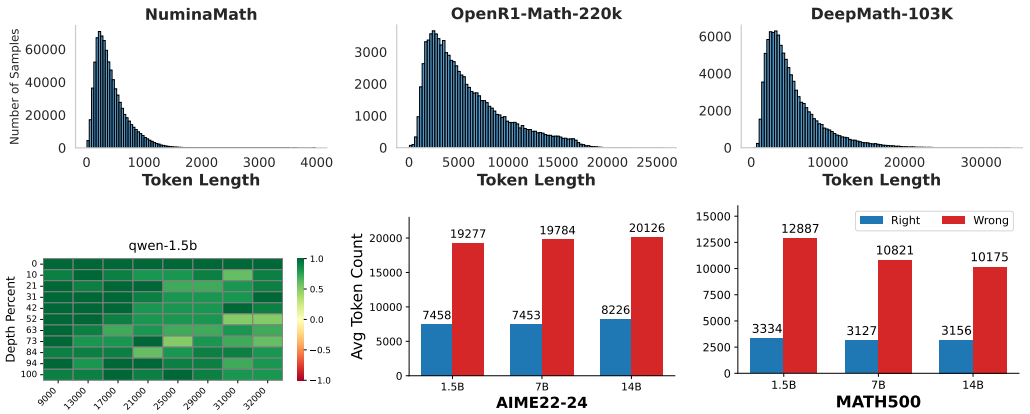


Figure 4: **Top:** Length distribution of three reasoning datasets. `NuminaMath-CoT` represents early chain-of-thought (CoT) data with short sequences, while `OpenR1-Math-220K` and `DeepMath-103K`, generated by `DeepSeek-R1`, exhibit significantly longer outputs. **Bottom-left:** Performance of `DeepSeek-Distilled-Qwen-1.5B` on the *Needle-in-a-Haystack* benchmark with 32K context. **Bottom-middle/right:** Average output lengths of correct and incorrect generations on `AIME` and `MATH500` for `DeepSeek-Distilled-Qwen-1.5B`, `7B`, and `14B`. Incorrect answers consistently exhibit longer output lengths, indicating potential limitations of long-context ability in reasoning.

To further investigate this hypothesis, we manually inspect a subset of long-output failures and identify two recurring patterns, repetition and contextual reference failures, due to limited long-context capability. The first pattern involves excessive repetition, where the model loops over the same sentence or phrase, failing to advance the solution, as shown in Figure 2. The second pattern arises in the latter part of the output, where the model incorrectly recalls earlier mathematical expressions from the problem, leading to flawed reasoning and incorrect conclusions, as shown in Figure 3. Both failure modes underscore the model's struggle to maintain coherence and accuracy over long sequences—a well-known limitation of inadequate long-context capacity that degrades and distorts in-context learning performance in LLMs.

2.3 Long and variable reasoning data necessitate long-context models.

Current reasoning models are typically fine-tuned on long chain-of-thought (CoT) datasets generated by large-scale reasoning models. A key characteristic of these models is their tendency to produce *long and variable-length outputs*. As a result, the resulting datasets exhibit broad length distributions, with many examples exceeding 10K tokens—far longer than early CoT-style data. This necessitates that fine-tuned models be capable of handling such long and diverse sequences. In Figure 4, we analyze the length distributions of three representative datasets: `NuminaMath-CoT`, `OpenR1-Math-220K`, and `DeepMath-103K`. `NuminaMath-CoT`, representing early CoT-style data, primarily contains samples under 1K tokens. In contrast, `OpenR1-Math-220K` and `DeepMath-103K`, both collected from `DeepSeek-R1`, contain significantly longer reasoning sequences, with a large proportion of samples exceeding 4K tokens and some even surpassing 10K tokens, which is totally different with early CoT datasets like `NuminaMath-CoT`.

There is a lack of research exploring how such increased sequence lengths interact with or depend on the model's long-context capability after the release of Deepseek-R1. For instance, if a model lacks sufficient long-context ability, training on long reasoning sequences may fail to yield expected improvements—or even negatively impact performance—due to the model's inability to effectively utilize the full input. While modern models are advertised to handle long sequences (e.g., `Qwen2.5-1.5B-Instruct` supports up to 32K tokens), their effective context length is often substantially shorter. As shown in Figure 4, we evaluate `Qwen2.5-1.5B-Instruct` on the *Needle-in-a-Haystack* benchmark and observe that the model fails to maintain high accuracy across all cases in 32k contexts, indicating its limitations in effective long-context processing and long-context ability.

Table 2: Effective context length and long-context benchmark performance of `LLaMA3-8B-Instruct` under different `RoPE theta` scaling factors. We report the estimated effective context length, Needle-in-a-Haystack (NIAH) retrieval accuracy at 32k, as well as performance on LongBench and RULER. Results show that scaling up to RoPE $\times 16$ consistently improves long-context robustness across all benchmarks, but further scaling (e.g., $\times 32$ and $\times 64$) leads to diminishing or even negative returns. Notably, the effective length surpasses the maximum sequence length (16k) of the current training dataset when the factor exceeds 4, which is sufficient for reasoning training.

RoPE theta	$\times 1$	$\times 4$	$\times 8$	$\times 16$	$\times 32$	$\times 64$
Effective Context Length	9k	21k	29k	32k	21k	17k
32k NIAH Score	0.00	3.75	58.30	77.05	58.86	35.00
LongBench Score	21.14	39.21	39.78	40.41	38.96	38.01
RULER Score	56.13	69.57	79.62	94.24	88.07	84.98

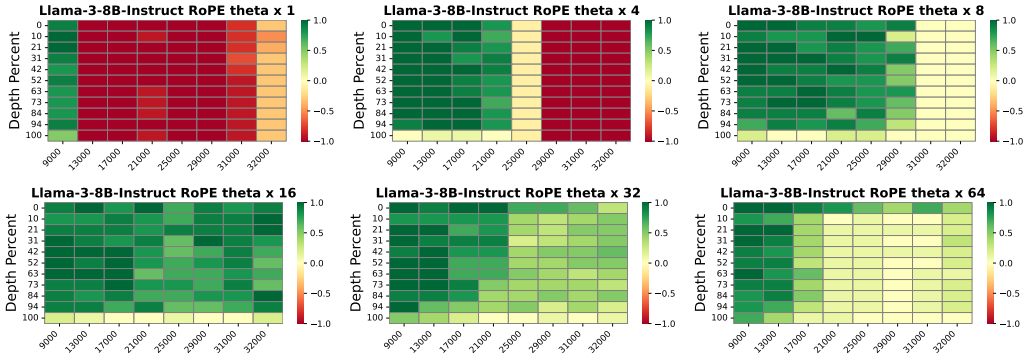


Figure 5: Needle-in-a-Haystack Results for LLaMA-3-8B-Instruct. Performance of LLaMA-3-8B-Instruct on the *Needle-in-a-Haystack* benchmark with a 32K context under different `RoPE theta` scaling factors. `RoPE theta x 16` refers to scaling the original `RoPE theta` by a factor of 16.

3 Empirical analysis: verifying the connection between reasoning and long-Context ability

In this section, we investigate the connection between a model's long-context capability and its reasoning performance. We begin by applying long-context extension strategies to obtain models with varying levels of long-context ability at 32k tokens. These models are then fine-tuned on different reasoning datasets using supervised fine-tuning (SFT), allowing us to assess how different context capacities influence the effectiveness of reasoning training. Next, we further extend the long-context capability to 128K tokens or beyond, in order to examine whether extreme context lengths can provide additional gains in reasoning performance. Finally, we introduce a recipe for LLM reasoning training: extend the long-context ability before reasoning finetuning and conduct an experiment to verify that this recipe is useful and effective. The overall training pipeline is illustrated in Figure 6.

3.1 Experimental setup

Long context ability extension strategy. We use two strategies to enhance long-context capability. The first is directly scaling the `RoPE theta` parameter by different factors, which has been shown to improve a model's ability to handle longer sequences [7]. The second leverages model merging: we merge the target model with another model that possesses stronger long-context capabilities. In this setting, we carefully control the merge ratio to ensure that the base performance remains nearly unchanged, allowing us to isolate the effect of long-context enhancement as the only influential factor.

Data processing for reasoning SFT. We utilize the OpenR1-Math-220K dataset [8] and divide it into two categories based on response length: **short** samples (responses within 8K tokens) and **long** samples (responses ranging from 8K to 16K tokens). For both categories, we sample 20K instances and perform correctness filtering to ensure that each response is factually accurate and correct. These two subsets are then used independently to fine-tune models to improve their reasoning ability.

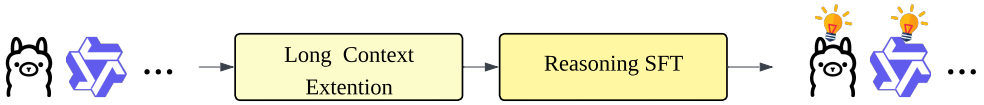
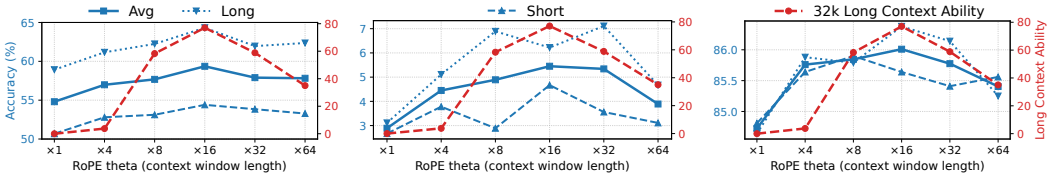


Figure 6: Pipeline for verifying the connection between reasoning and long-context ability. We first expand the model's long-context capability to obtain variants with different levels of long-context capacity. Then, we perform SFT on reasoning datasets to obtain reasoning-enhanced models.

Training details. All models are fine-tuned using four NVIDIA H200 GPUs. We employ the `LLaMAFactory` library with a batch size of 32, a learning rate of 1.0×10^{-5} and 3 epochs.

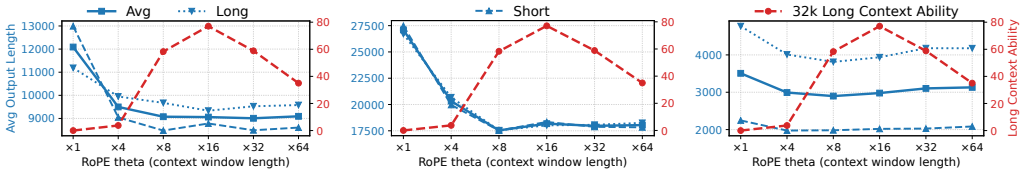
Long context evaluation. We adopt the *Needle in a Haystack* benchmark provided by the OpenCompass framework to access long context ability. For simplicity and robustness, we use the accuracy on the single-haystack setting as our primary metric: a correct response receives a score of $+1$, a repetitive/degenerate answer receives -1, and an incorrect but non-degenerate response receives 0.

Reasoning evaluation. To further evaluate the model's reasoning ability post-training, we use three math benchmarks: MATH500, AIME22–24, and GSM8K. Following the evaluation methodology from DeepSeek-R1, we adopt the *pass@1(5)* metric, where five responses are generated for each question and accuracy is computed over all the responses. This provides a more stable estimate of reasoning performance and abilities of Large Language Models after finetuning with datasets.



(a) MATH500 (b) AIME22-24 (c) GSM8K

Figure 7: Visualization of the relationship between 32k long-context ability and reasoning performance across different model variants of `LLaMA-3.1-8B-Instruct` on three math benchmarks (MATH500, AIME22-24, GSM8K). `Short` and `Long` refer to performance after fine-tuning on short and long reasoning datasets. `Avg` represents the average of the short and long fine-tuned results.



(a) MATH500 (b) AIME22-24 (c) GSM8K

Figure 8: Visualization of the relationship between 32k long-context ability and average output length across different model variants of `LLaMA-3.1-8B-Instruct` on MATH500, AIME22-24, GSM8K.

3.2 How does long-context capacity affect reasoning SFT?

In this experiment, we investigate how different levels of long-context capability affect the performance of reasoning SFT (Supervised Finetuing). We use `LLaMA3-8B-Instruct` as the base model, which originally supports up to 8K tokens. To obtain models with varying long-context capabilities, we scale the `RoPE theta` value by the following different factors: {1, 4, 8, 16, 32, 64}.

First, we evaluate the long-context ability of each variant using the *Needle-in-a-Haystack* benchmark. As shown in Figure 5 and Table 2, performance improves as the `RoPE theta` scaling factor increases from 1 to 16, reaching peak performance at 16. Beyond this point, further increases in `theta` lead to performance degradation. What's more, when the scaling theta is more than 4, the model's effective length is more than 16k, which is enough for the reasoning training for current datasets.

Table 3: **Effect of RoPE theta scaling on long-context and reasoning performance on MATH500.** Evaluation of `LLaMA3-8B-Instruct` with different `RoPE theta` scaling factors on the 32K *Needle-in-a-Haystack* benchmark. `Base` refers to the model's performance before SFT, while `Short` and `Long` denote results after fine-tuning on short and long reasoning datasets, respectively. `Avg` represents the average of the `Short` and `Long` performances.

RoPE	32k long	Acc(%)				Average Output Length			
theta	ctx ability	Base	Short	Long	Avg	Base	Short	Long	Avg
×1	0	**24.40**	50.68	58.92	54.80	746	12991	11187	12089
×4	3.75	20.40	52.80	61.16	56.98	754	9047	9949	9498
×8	58.30	18.96	53.12	62.24	57.68	767	**8476**	9673	9075
×16	**77.05**	17.28	**54.40**	**64.32**	**59.36**	608	8782	**9335**	9059
×32	58.86	15.24	53.84	61.96	57.90	490	8489	9526	**9007**
×64	35.00	14.20	53.28	62.36	57.82	453	8605	9579	9092

Table 4: Cross-domain evaluation of long-context ability on GPQA (science) and Livecode (code).

Task	Setting	GPQA Accuracy (%)			Livecode Accuracy (%)		
		×1	×4	×16	×1	×4	×16
32k NIAH	-	0.00	3.75	77.05	0.00	3.75	77.05
Accuracy	before training	31.82	31.31	30.81	12.30	10.16	7.45
	after training	37.27	39.19	41.92	25.34	27.95	32.27

Next, we perform supervised fine-tuning (SFT) using both `short` and `long` reasoning datasets on models with different `RoPE theta` settings. The detailed results are summarized in Table 3. In addition, we plot the relationship between each model's long-context capability and its reasoning performance. We also visualize how long-context ability correlates with the average output length. The results are shown in Figures 7 and 8. We observe that the best reasoning performance across all three evaluation benchmarks consistently occurs when the `RoPE theta` scaling factor is set to 16—coinciding with the strongest long-context capability observed in the Needle-in-a-Haystack task. In contrast, models with suboptimal `theta` settings exhibit reduced accuracy undergoing the same SFT procedure. These findings suggest that enhancing long-context processing ability can directly contribute to improved reasoning performance after reasoning SFT (Supervised Fine-tuning).

Furthermore, we observe that, in general, models fine-tuned on `long` reasoning datasets outperform those fine-tuned on `short` datasets in both `MATH500` and `AIME`. These long-form datasets typically contain more complex problems and richer intermediate reasoning steps, making them more effective for improving the model's reasoning ability. However, in order to benefit from such data, models must possess sufficiently strong and effective long-context capabilities to process the extended inputs.

3.3 Generality and robustness of long-context gains to reasoning ability

We investigate whether the benefits of stronger long-context ability generalize across tasks, model families, and input lengths. Our results show that the improvements are not limited to math reasoning.

Cross-domain generalization. We fine-tuned `LLaMA3-8B-Instruct` models with RoPE scales ×1, ×4, and ×16 on 20k samples from the *science* and *code* domains of `OpenThoughts3-1.2M`. Evaluations on GPQA and Livecode (Table 4) show that stronger long-context ability consistently improves accuracy beyond math reasoning.

Across model families. We trained `Phi-4` (14B) with the same RoPE settings. Consistent gains on MATH500 (Table 5) confirm that the benefit is not tied to a specific architecture or scale. The results of AIME22-24 are in Table 10 and it has the similar performance like MATH500.

Short-input reasoning. On the short-input `MMLU-STEM` benchmark, models with stronger long-context ability also achieve higher post-training accuracy (Table 6). This indicates that long-context training not only preserves but can reinforce reasoning on short-input tasks.

Table 5: Performance of `Phi-4` under different RoPE scales on MATH500.

Model	RoPE	32k NIAH (%)	MATH500 (before, %)	MATH500 (after, %)
Phi-4	×1	52.27	79.52	88.62
Phi-4	×4	78.07	77.48	89.14
Phi-4	×16	84.77	73.20	89.90

Table 6: Effect of long-context ability on short-input reasoning (MMLU-STEM) with `LLaMA3-8B`.

RoPE Scale	32k NIAH (%)	MMLU-STEM (before, %)	MMLU-STEM (after, %)
×1	0.00	54.36	71.06
×4	3.75	53.85	73.04
×16	77.05	51.44	74.27

Overall, these results demonstrate that the gains from stronger long-context ability are robust: they extend beyond math to science and code, hold across LLaMA, Qwen, and Phi model families, and even benefit short-input reasoning.

3.4 Does extremely long context bring further gains on reasoning SFT?

In previous experiments, we extended models to handle up to 32K tokens and observed notable improvements in reasoning performance. A natural question arises: *can even stronger long-context capabilities further enhance reasoning*, or is there a limit beyond which performance saturates or even degrades? To explore this, we conduct experiments using models with a context length of 1M tokens (Qwen2.5-7B-Instruct-1M)—far beyond the typical range of existing reasoning datasets.

We adopt a linear merging strategy to construct models with varying long-context capacities while minimizing changes to their base ability. Specifically, we merge two models with different context capabilities at various ratios to obtain models with intermediate long-context strengths. With selected merge ratio, we ensure that the base capabilities remain largely unchanged, isolating long-context capacity as the key variable. We apply this strategy to two models: Qwen2.5-7B-Instruct-1M with long context ability at 1M and Qwen2.5-7B-Instruct with long context ability at 32k.

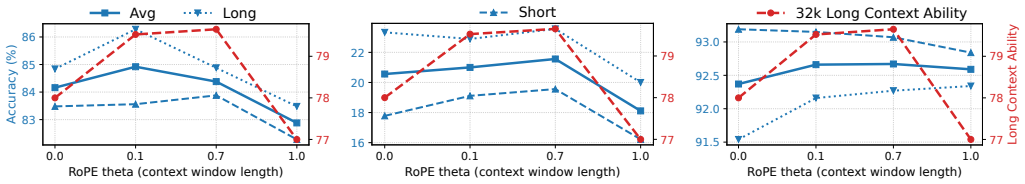
We first evaluate the merged models on *Needle-in-a-Haystack* at 32k and find that long-context ability grows with the merge ratio, but near-perfect scores make it less discriminative (Appendix A). We therefore adopt more challenging 32k tasks such as *Value Tracking* and *Question Answering*, which better capture effective long-context processing by requiring stronger long-range reasoning.

We fine-tune the merged models on both `short` and `long` reasoning datasets, and evaluate on `MATH500`, `AIME22-24`, and `GSM8K`. We also examine the relationship between effective long-context ability, reasoning accuracy, and output length (Figures 9 and 10). Results show that moderate merge ratios (e.g., 0.1, 0.7) yield strong effective long-context ability and high reasoning accuracy, while the 1M model (ratio 1.0) exhibits weaker effective long-context utilization and degraded performance.

3.5 Proposed reasoning recipe: extend context length first

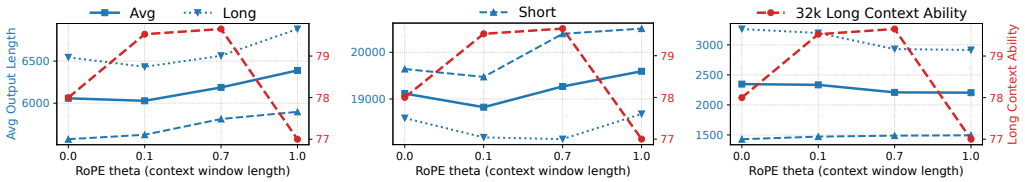
Based on our previous experiments, we propose a training **recipe** for improving reasoning capabilities via supervised fine-tuning (SFT): *first appropriately enhance the model's long-context capability, then apply reasoning-specific SFT*. This recipe aims to better prepare the model for long-form reasoning tasks, where both context length and reasoning complexity are critical. To validate this approach, we experiment with the `Qwen2.5-Math-7B-Instruct` model, which demonstrates strong mathematical performances but has a limited context length of 4k tokens, which is a good example of the recipe.

We enhance its long-context capability by multiplying the `RoPE theta` value by a scaling factor of 16 and merging the model with `Qwen2.5-7B-Instruct-1M` using a merge ratio of 0.3. After the merging step, the long-context ability improves noticeably, as is shown in Appendix A. Subsequently, we perform SFT using both `short` and `long` reasoning datasets, and evaluate the fine-tuned models on two benchmarks: `MATH500` and `AIME22-24`. As shown in Table 7, across both short and long



(a) MATH500 (b) AIME22-24 (c) GSM8K

Figure 9: Visualization of the relationship between 32k long-context ability and reasoning performance across different model variants of `Qwen2.5-7B-Instruct` on MATH500, AIME22-24, GSM8K. A merge ratio of 0.1 indicates that the long-context variant (`Qwen2.5-7B-Instruct-1M`) contributes 10% to the final merged model. `Short` and `Long` refer to performance after fine-tuning on short and long reasoning datasets. `Avg` represents the average of the short and long fine-tuned results.



(a) MATH500 (b) AIME22-24 (c) GSM8K

Figure 10: The relationship between 32k long-context ability and average output length across different model variants of `Qwen2.5-7B-Instruct` on three math benchmarks.

reasoning datasets, the model variant with `RoPE theta` scaled by 16 and a merge ratio of 0.3 consistently achieves the highest performance after reasoning SFT on short and long datasets.

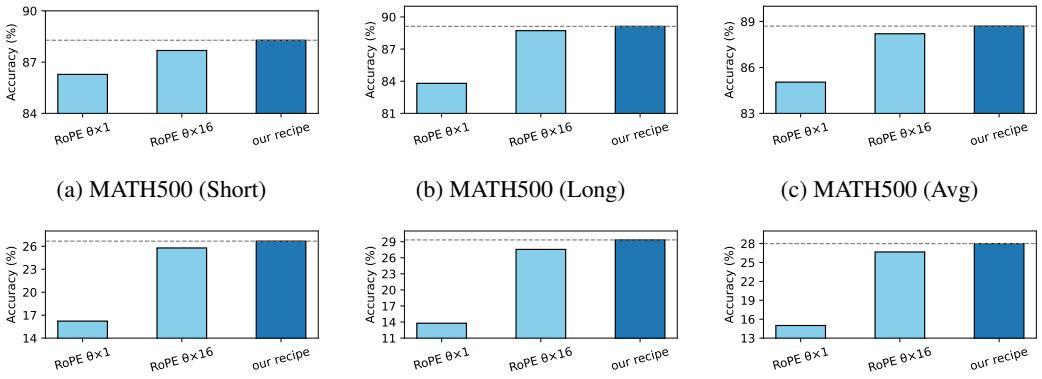
4 Related Works

Reasoning models. The release of `DeepSeek-R1` [9] has catalyzed widespread interest in open-source reasoning models [10–12]. Most existing approaches to improving reasoning rely on supervised fine-tuning (SFT) and rule-based reinforcement learning (RL) [13–16]. For instance, `SkyThought` [17] and `Bespoke` [18] fine-tune models using responses generated by larger reasoning models such as `QwQ-32B` [19] and `DeepSeek-R1`. Other methods like `S1` [20] and `LIMO` [21] highlight the importance of using compact yet high-quality reasoning samples to improve data efficiency. These efforts have led to the creation of numerous open-source reasoning datasets, including `OpenR1-Math-220K` [8], `DeepMath-103K` [4], `OpenThoughts` [5], and others [6, 22, 23]. Notable models trained on these datasets include `OpenR1-Qwen-7B` and `DeepMath-Zero-7B`. Subsequent work [24, 25] has proposed techniques to further optimize SFT and RL pipelines, arguing that structural aspects of reasoning data (e.g., step-wise decomposition) can be more impactful than content alone. More recent studies [3, 26–28] aim to improve the alignment and generalization of reasoning through better data construction, gradient control, and training optimization strategies. Despite these advances, an important factor remains underexplored: the role of `long-context capability` in reasoning. While recent datasets and tasks increasingly involve longer sequences, no existing work systematically investigates how a model's ability to process extended contexts affects its reasoning performance. Our work aims to figure out the relationship between them.

Long-context ability. Closed-source models such as `GPT-4` [29], `Claude` [30], and Gemini [31] already support context lengths of 128K tokens or more, enabled by advances in both pre-training and post-training [32–35]. In parallel, open-source models—including `LLaMA` [36], `Qwen` [37, 38], and `Phi` [39]—have also extended their capacities, with some reaching 1M tokens, such as `Qwen-2.5-7B-Instruct-1M` [40] and `LLaMA-3.1-Nemotron-8B-UltraLong-4M-Instruct` [41]. To enable long context, several approaches have been proposed. *RoPE-based methods* adapt positional encoding for extrapolation without full retraining, including Position Interpolation [42], NTK Scaling [7], YaRN [43], and SelfExtend [44]. *Attention redesigns* improve scalability or memory retention, such as StreamingLLM [45], LM-Infinite [46], Inf-LLM [47], and Landmark Attention [48]. Another direction is *input compres-*

Table 7: **Validation of the proposed fine-tuning recipe for reasoning.** We evaluate the effectiveness of our proposed training recipe using `Qwen2.5-Math-7B-Instruct`, a model with an initial long-context capacity limited to 4K tokens. We first scale its `RoPE theta` by a factor of 16, and then merge it with `Qwen2.5-7B-Instruct-1M` using a merge ratio of 0.3. We observe consistent improvement in reasoning performance after each step, demonstrating the effectiveness of enhancing long-context capability prior to reasoning fine-tuning to get a higher reasoning ability.

Operation	Acc(%)				Avg Output Length			
	Base	Short	Long	Avg	Base	Short	Long	Avg
MATH500								
RoPE theta $\times 1$	**81.88**	86.28	83.80	85.04	2037	1022	2147	1584
RoPE theta $\times 16$	65.16	87.68	88.72	88.20	4579	4213	5789	5501
$0.7 \times$ RoPE theta $\times 16 + 0.3 \times$ 1M-Model	74.12	**88.28**	**89.12**	**88.70**	2228	**4948**	**6515**	**5731**
AIME22-24								
RoPE theta $\times 1$	**8.44**	16.22	13.78	15.00	8257	3463	10418	6940
RoPE theta $\times 16$	3.78	25.78	27.56	26.67	13411	15036	17321	16179
$0.7 \times$ RoPE theta $\times 16 + 0.3 \times$ 1M-Model	7.11	**26.67**	**29.33**	**28.00**	6859	**20265**	**22919**	**21592**



(a) MATH500 (Short) (b) MATH500 (Long) (c) MATH500 (Avg)

(d) AIME22-24 (Short) (e) AIME22-24 (Long) (f) AIME22-24 (Avg)

Figure 11: Comparison of reasoning accuracy across different configurations. Results are reported on `MATH500` (top row) and `AIME22-24` (bottom row), with separate evaluation for short, long, and average context cases. Our proposed recipe (dark blue bar, dashed line) consistently outperforms the baseline `RoPE theta` $\times 1$ and the extended-context baseline `RoPE theta` $\times 16$, demonstrating that enhancing long-context capability prior to reasoning fine-tuning yields stable and significant accuracy improvements, enhancing models' reasoning ability.

sion, which reduces sequence length by summarization or filtering [49, 50]. Finally, recent work explores hardware-aligned designs, including sparse attention architectures [51] and MoBA [52].

5 Discussion and Conclusions

Limitations: Our analysis is limited to 7B–8B models and does not cover larger scales. We also focus on supervised fine-tuning, leaving open how reinforcement learning interacts with long-context ability. Future work should examine these directions across model families and scales.

This paper investigates the overlooked yet crucial role of long-context capability in reasoning performance. Through behavioral analysis and controlled experiments, we demonstrate a consistent correlation between long-context and downstream reasoning ability. Our results show that models with improved long-context capacity not only perform better on tasks involving lengthy inputs but also achieve higher reasoning accuracy even on short-form benchmarks. Furthermore, we find that enhancing long-context ability prior to supervised fine-tuning yields significant gains across multiple reasoning datasets. Based on these findings, we advocate for a training recipe: first, extend the model's long-context ability, then apply reasoning-specific fine-tuning.

Acknowledgments

This work was supported in part by NSF grants 2112606 and 2117439. Further, this research made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University (CWRU). We give special thanks to the CWRU HPC team for their prompt and professional help and maintenance.

References

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

[2] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[3] Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report, 2025.

[4] Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning, 2025.

[5] OpenThoughts Team. Open Thoughts. https://open-thoughts.ai, January 2025.

[6] Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025.

[7] 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.

[8] Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025.

[9] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, and Ruoyu Zhang... Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[10] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*, 2024.

[11] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024.

[12] Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, et al. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*, 2025.

[13] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[14] Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922*, 2024.

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[16] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025.

[17] NovaSky Team. Sky-t1: Train your own o1 preview model within $450. https://novasky-ai.github.io/posts/sky-t1, 2025. Accessed: 2025-01-09.

[18] Bespoke Labs. Bespoke-stratos: The unreasonable effectiveness of reasoning distillation. www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoning-distillation, 2025. Accessed: 2025-01-22.

[19] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025.

[20] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025.

[21] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025.

[22] Sathwik Tejaswi Madhusudhan, Shruthan Radhakrishna, Jash Mehta, and Toby Liang. Millions scale dataset distilled from r1-32b. https://huggingface.co/datasets/ServiceNow-AI/R1-Distill-SFT, 2025.

[23] Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, and Ehud Karpas.. Llama-nemotron: Efficient reasoning models, 2025.

[24] Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Llms can easily learn to reason from demonstrations structure, not content, is what matters!, 2025.

[25] Ke Ji, Jiahao Xu, Tian Liang, Qiuzhi Liu, Zhiwei He, Xingyu Chen, Xiaoyuan Liu, Zhijie Wang, Junying Chen, Benyou Wang, et al. The first few tokens are all you need: An efficient and effective unsupervised prefix fine-tuning method for reasoning models. *arXiv preprint arXiv:2503.02875*, 2025.

[26] Jiarui Yao, Yifan Hao, Hanning Zhang, Hanze Dong, Wei Xiong, Nan Jiang, and Tong Zhang. Optimizing chain-of-thought reasoners via gradient variance minimization in rejection sampling and rl, 2025.

[27] Tianwei Ni, Allen Nie, Sapana Chaudhary, Yao Liu, Huzefa Rangwala, and Rasool Fakoor. Teaching large language models to reason through learning and forgetting, 2025.

[28] Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025.

[29] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[30] Loredana Caruccio, Stefano Cirillo, Giuseppe Polese, Giandomenico Solimando, Shanmugam Sundaramurthy, and Genoveffa Tortora. Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach. *Intelligent Systems with Applications*, 21:200336, 2024.

[31] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[32] 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.

[33] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*, 2023.

[34] Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, and Yanning Chen. Liger kernel: Efficient triton kernels for llm training. *arXiv preprint arXiv:2410.10989*, 2024.

[35] Shenggui Li, Fuzhao Xue, Chaitanya Baranwal, Yongbin Li, and Yang You. Sequence parallelism: Long sequence training from system perspective. *arXiv preprint arXiv:2105.13120*, 2021.

[36] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, and Angela Fan... The llama 3 herd of models, 2024.

[37] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[38] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, and Dayiheng Liu ... Qwen3 technical report, 2025.

[39] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024.

[40] An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. Qwen2.5-1m technical report, 2025.

[41] Chejian Xu, Wei Ping, Peng Xu, Zihan Liu, Boxin Wang, Mohammad Shoeybi, and Bryan Catanzaro. From 128k to 4m: Efficient training of ultra-long context large language models. *arXiv preprint*, 2025.

[42] 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.

[43] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

[44] Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. Llm maybe longlm: Self-extend llm context window without tuning, 2024.

[45] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[46] 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.

[47] Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. Infllm: Training-free long-context extrapolation for llms with an efficient context memory. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[48] Amirkeivan Mohtashami and Martin Jaggi. Landmark attention: Random-access infinite context length for transformers. *arXiv preprint arXiv:2305.16300*, 2023.

[49] Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*, 2023.

[50] Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. Compressing context to enhance inference efficiency of large language models. *arXiv preprint arXiv:2310.06201*, 2023.

[51] Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Y. X. Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native sparse attention: Hardware-aligned and natively trainable sparse attention, 2025.

[52] Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, Zhiqi Huang, Huan Yuan, Suting Xu, Xinran Xu, Guokun Lai, Yanru Chen, Huabin Zheng, Junjie Yan, Jianlin Su, Yuxin Wu, Yutao Zhang, Zhilin Yang, Xinyu Zhou, Mingxing Zhang, and Jiezhong Qiu. Moba: Mixture of block attention for long-context llms. *arXiv preprint arXiv:2502.13189*, 2025.

[53] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models, 2025.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: yes, we include the paper's contributions and scope in the abstract and introduction

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: we add the limitations to the conclusion

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: we include the full set of assumptions and a complete (and correct) proof

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we include a section about the experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: we provide open access to the data and code

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: we include a section about the experimental setup

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: we include a stable metric to get results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification:: we provide sufficient information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: the research conducted in the paper conform with the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: there is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: the paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: the paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: we do not involve LLMs to impact the methodology.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Other Results on Different Long-Context Ability

This section provides additional experimental results that complement the main findings by examining how different levels of long-context ability influence model performance across multiple reasoning benchmarks. While the main paper focused on reasoning accuracy trends, here we provide a broader view including auxiliary benchmarks, output length analysis, and merged-model experiments.

Effect of RoPE scaling. We first analyze the impact of varying `RoPE` θ scaling factors on `LLaMA3-8B-Instruct`. As shown in Table 8, models with stronger long-context ability generally achieve higher reasoning accuracy across `MATH500`, `AIME22-24`, and `GSM8K`. The trend is most consistent up to RoPE $\times 16$, after which performance begins to plateau or slightly decline, suggesting diminishing returns when scaling beyond the effective context length of the training data. We also observe that output length tends to shrink as accuracy improves, consistent with our analysis in Section 3.2, where correct solutions are typically shorter and more concise.

Table 8: **Effect of RoPE θ Scaling on Long-Context and Reasoning Performance.** Evaluation of `LLaMA3-8B-Instruct` with different `RoPE` θ scaling factors on the 32K *Needle-in-a-Haystack* benchmark. `Base` refers to the model's performance before SFT, while `Short` and `Long` denote results after fine-tuning on short and long reasoning datasets, respectively. `Avg` represents the average of the `Short` and `Long` performances. Length represents the average output length.

RoPE	32k long	Acc(%)				Length			
theta	ctx ability	Base	Short	Long	Avg	Base	Short	Long	Avg
MATH500									
$\times 1$	0	**24.40**	50.68	58.92	54.80	746	12991	11187	12089
$\times 4$	3.75	20.40	52.80	61.16	56.98	754	9047	9949	9498
$\times 8$	58.30	18.96	53.12	62.24	57.68	767	**8476**	9673	9075
$\times 16$	**77.05**	17.28	**54.40**	**64.32**	**59.36**	608	8782	9335	**9059**
$\times 32$	58.86	15.24	53.84	61.96	57.90	490	8489	**9526**	9007
$\times 64$	35.00	14.20	53.28	62.36	57.82	453	8605	9579	9092
AIME22-24									
$\times 1$	0	0.22	2.67	3.11	2.89	1895	27482	26706	27094
$\times 4$	3.75	**0.67**	3.78	5.11	4.45	2048	19961	20684	20322
$\times 8$	58.30	0.22	2.89	6.89	4.89	1723	**17532**	**17547**	**17539**
$\times 16$	**77.05**	0.22	**4.67**	6.22	**5.45**	1617	18339	18015	18177
$\times 32$	58.86	0.00	3.56	**7.11**	5.34	1372	17876	18071	17973
$\times 64$	35.00	0.44	3.11	4.67	3.89	1108	17830	18251	18041
GSM8K									
$\times 1$	0	**78.13**	84.81	84.64	84.73	195	2244	4769	3506
$\times 4$	3.75	75.54	85.64	85.88	85.76	190	**1973**	4010	2992
$\times 8$	58.30	73.03	**85.90**	85.79	85.85	184	1981	**3812**	**2897**
$\times 16$	**77.05**	70.39	85.64	**86.38**	**86.01**	182	2019	3936	2977
$\times 32$	58.86	66.35	85.41	86.14	85.78	180	2026	4179	3103
$\times 64$	35.00	62.21	85.56	85.25	85.41	182	2083	4177	3130

Extremely long context via model merging. To test whether extremely long contexts (e.g., 128k or 1M) provide further gains, we construct merged variants of `Qwen2.5-7B-Instruct` by combining its base 32k model with the ultra-long version `Qwen2.5-7B-Instruct-1M`. By adjusting merge ratios, we obtain intermediate models with controllable long-context capabilities, while preserving base reasoning ability. As shown in Figure 9, reasoning performance correlates with effective long-context strength: models with moderate merge ratios (e.g., 0.1, 0.7) achieve consistently strong accuracy, whereas the pure 1M model shows weaker effective long-context utilization and degraded reasoning.

B Other Results on Different Merging Ratios

We further analyze how different merging ratios between base and ultra-long variants affect both retrieval ability and downstream reasoning. This provides insight into whether extremely long contexts (e.g., 1M tokens) always yield benefits, or whether moderate ratios strike a better balance between long-context integration and base reasoning stability.

Table 9: **Reasoning Performance After SFT with Different Merge Ratios.** Qwen results are based on merging `Qwen2.5-7B-Instruct-1M` with `Qwen2.5-7B-Instruct`. A ratio of 0.1 means that the long-context (1M) variant contributes 10% to the merged model.

1M Merge	32k long	Acc(%)				Length			
Ratio	ctx ability	Base	Short	Long	Avg	Base	Short	Long	Avg
MATH500									
0	78.1	**75.00**	83.48	84.84	84.16	638	5572	6545	6058
0.1	79.1	74.64	83.56	**86.28**	**84.92**	670	5625	6432	**6028**
0.7	79.5	72.80	**83.88**	84.88	84.38	630	5812	6563	6187
1.0	77.7	72.16	82.28	83.48	82.88	688	5897	6881	6389
AIME22-24									
0	78.1	8.22	17.78	23.33	20.56	1051	19642	18592	19117
0.1	79.1	**9.33**	19.11	22.89	21.00	1313	19473	18178	**18825**
0.7	79.5	7.78	**19.56**	**23.56**	**21.56**	1416	20396	18142	19269
1.0	77.7	7.33	16.22	20.00	18.11	1750	20507	18680	19594
GSM8K									
0	78.1	**90.75**	**93.19**	91.54	92.37	259	1429	3264	2347
0.1	79.1	**90.75**	93.15	92.16	92.66	252	1472	3197	2334
0.7	79.5	89.1	93.07	**92.27**	**92.67**	254	1488	2930	2209
1.0	77.7	89.13	92.84	92.34	92.59	254	1494	2915	**2204**

Needle-in-a-Haystack (retrieval ability). We first evaluate the merged models on the 32K *Needle-in-a-Haystack* benchmark. As shown in Figure 12, long-context ability generally increases with higher contribution from the 1M-token variant. However, because most models achieve near-perfect scores, the benchmark does not fully differentiate their effective long-context strength. This motivates evaluating more challenging reasoning datasets where differences manifest more clearly.

Reasoning benchmarks. We then fine-tune the merged models on both short and long reasoning datasets and evaluate them on three benchmarks: `MATH500`, `AIME22-24`, and `GSM8K`. Table 9 reports accuracy and output length. Three key observations emerge: (1) Moderate merge ratios (e.g., 0.1 or 0.7) yield the best overall reasoning performance, consistently outperforming both the base (0) and fully merged (1.0) variants. (2) Pure 1M models (ratio 1.0) exhibit degraded reasoning accuracy, despite higher nominal long-context capacity, suggesting weaker effective utilization. (3) Output lengths tend to shrink as reasoning improves, consistent with the observation that successful reasoning requires fewer redundant tokens.

C Additional Analysis on Output Length and Context Extension

Why is the output length decreasing with the long context extension? The observed decrease in output length is actually aligned with the ratio of correct answers. Correct generations tend to be much shorter and more concise than incorrect ones. With long-context extension, the number of correct answers increases, thus reducing the overall average output length. This phenomenon has also been discussed in prior work [53].

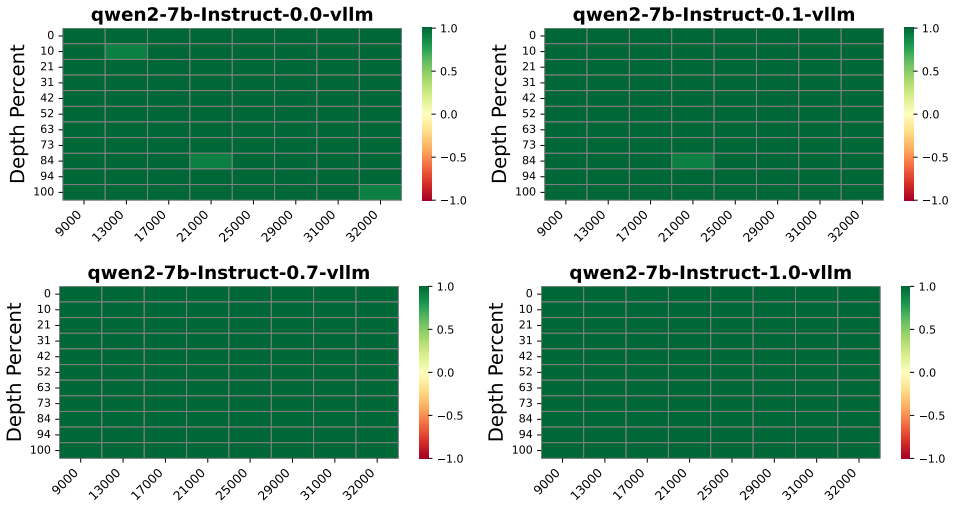


Figure 12: 32K Needle-in-a-Haystack evaluation for different merge ratios. Models with higher long-context contribution achieve stronger retrieval ability, but performance differences saturate quickly.

Table 10: Performance of `Phi-4` under different RoPE scales on AIME.

Model	RoPE	32k NIAH (%)	AIME (before, %)	AIME (after, %)
Phi-4	$\times 1$	52.27	14.89	47.56
Phi-4	$\times 4$	78.07	13.78	49.45
Phi-4	$\times 16$	84.77	10.22	50.17

To support this explanation, we report the average lengths of correct and incorrect generations across both short (0–8k) and long (8–16k) training settings for `LLaMA3-8B` under different RoPE configurations: *RoPE* $\times 1$ *(8k context)* and *RoPE* $\times 16$ *(Extended context)*

Table 11: Effect of training length on output length under different RoPE scales.

RoPE Scale	Setting	Accuracy (%)	Avg Length	# Correct	Avg	# Wrong	Avg
$\times 1$	short	50.68	12991	1267	3189	1233	23145
$\times 1$	long	58.92	11187	1473	5928	1027	20999
$\times 16$	short	54.40	8782	1360	3906	1140	14921
$\times 16$	long	64.32	9335	1608	6063	892	17265

As shown, models with longer context capacity (e.g., RoPE $\times 16$) achieve higher accuracy while generating shorter outputs on average—primarily because a larger portion of the outputs are correct, and correct answers are generally shorter.