Chain-of-Thought Matters: Improving Long-Context Language Models with Reasoning Path Supervision

Dawei Zhu [*] [♡♣]   **Xiyu Wei** [*] [♠♣]   **Guangxiang Zhao** [†] [◇]   **Wenhao Wu** [♡♣]
Haosheng Zou [△] **Junfeng Ran** [♠♣]   **Xun Wang** [♡♣]
Lin Sun [‡] [◇]   **Xiangzheng Zhang** [‡] [◇]   **Sujian Li** [‡] [♡♣]

[♡] School of Computer Science, Peking University
[♠] School of Software and Microelectronics, Peking University
[♣] Key Laboratory of Computational Linguistics, Ministry of Education, Peking University
[◇] Qiyuan Tech   [△] ByteDance

Abstract

Recent advances in Large Language Models (LLMs) have highlighted the challenge of handling long-context tasks, where models need to reason over extensive input contexts to aggregate target information. While Chain-of-Thought (CoT) prompting has shown promise for multi-step reasoning, its effectiveness for long-context scenarios remains underexplored. Through systematic investigation across diverse tasks, we demonstrate that CoT's benefits generalize across most long-context scenarios and amplify with increasing context length. Motivated by this, we propose a process-supervised framework that teaches models to generate high-quality reasoning paths for enhanced long-context performance. Our framework incorporates a self-sampling mechanism to bootstrap reasoning paths and a novel quality assessment protocol specifically designed for long-context scenarios. Experimental results on various long-context benchmarks demonstrate the effectiveness of our approach, achieving significant improvements over outcome supervision baselines on both in-domain tasks (+13.6,+3.8 and +7.5 points for LLaMA/Qwen on MuSiQue) and cross-domain generalization (+9.3,+8.1 and +10.4 points on average across diverse QA tasks). We release our code and dataset at `https://github.com/lemon-prog123/LongRePS`.

1 Introduction

Large Language Models (LLMs) have revolutionized language modeling and achieved remarkable success in traditional NLP tasks (Brown et al., 2020). This success has spurred their application to increasingly complex real-world scenarios (Guo et al., 2025; Claude, 2024; GPT4o, 2024), particularly long-context tasks such as document-level
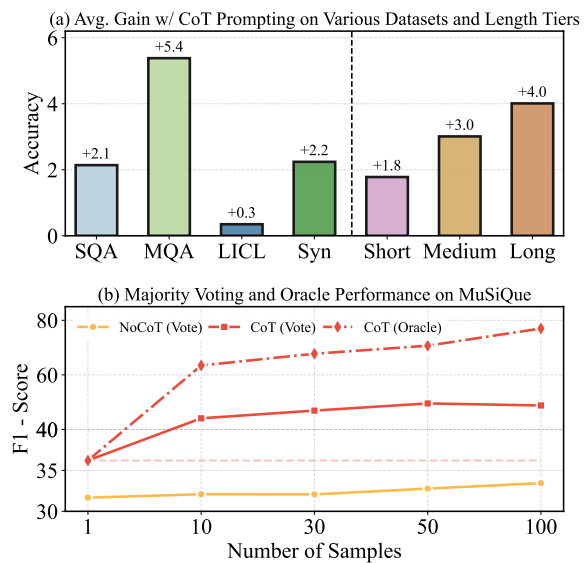

Figure 1: **(a)** Average gain w/ CoT prompting of open-source and proprietary models on long-context datasets of various domains and length tiers. *SQA*, *MQA*, *LICL*, *Syn* is short for *Single-Document QA*, *Multi-Document QA*, *Long In-Context Learning*, and *Synthetic* tasks, respectively. *Short*, *Medium*, *Long* denotes different length tiers (<32k, 32-96k, >96k). Details see Section 3. **(b)** Zero-shot majority voting results w.r.t. sampling rounds on MuSiQue, w/ and w/o CoT prompting.

question answering (Trivedi et al., 2022), summarization (Zhong et al., 2021), multi-shot in-context learning (Li et al., 2024b), and repository-level code generation (Bogomolov et al., 2024). These tasks typically require implicit reasoning steps that retrieve and aggregate information dispersed throughout extensive contexts before generating responses, posing significant challenges for contemporary long-context language models (LCLMs).

As a means to elicit reasoning, Chain-of-Thought (CoT) prompting has demonstrated remarkable efficacy in enhancing multi-step reasoning tasks (Wei et al., 2022; Madaan and Yazdanbakhsh, 2022; Sprague et al., 2024). However, its

---

[*]Equal contribution.
[†]Primary mentor
[‡]Corresponding authors.

effectiveness in long-context scenarios remains underexplored and only a few works have analyzed the use of CoT from limited perspectives, such as context length (Modarressi et al., 2025), task difficulty (Bai et al., 2024b), etc. Building on prior work, we conduct first systematic investigation of CoT's effectiveness across a diverse set of long-context tasks of varying lengths and domains on both open-source and proprietary models of different scales. As illustrated in Figure 1(a), CoT's benefits generalize across most long-context scenarios, and amplify with increasing context length. Even more intriguingly, zero-shot majority voting results (Figure 1) on the long-context multi-hop reasoning task MuSiQue (Trivedi et al., 2022) demonstrate that CoT substantially outperforms pure majority voting, and exhibits a steeper improvement trajectory as sampling rounds increase, but still significantly lags behind its oracle performance. This naturally leads to the following research question (**RQ**): *How to enhance models' capability in long-context scenarios to generate high-quality reasoning paths* [1] *for improved performance?*

To address the question, we further investigate methods to train models in generating high-quality reasoning paths in long-context scenarios, aiming to enhance overall performance. Given that real-world annotated data typically only contains outcome labels, we construct training data through a two-stage approach: first bootstrapping reasoning paths from the model itself, then identifying high-quality ones via a novel assessment protocol specifically designed for long-context scenarios. Our protocol evaluates both *answer correctness* and *process reliability*, with the latter strategically decomposed into *source faithfulness* and *intrinsic consistency* for efficient and accurate assessment. These self-sampling and quality assessment mechanisms together form a process-supervised framework that not only guides models toward correct answers but also teaches them appropriate reasoning patterns to reach these conclusions.

We train LLaMA3.1-8B (Dubey et al., 2024) and Qwen2.5-{7B,32B} (Yang et al., 2024a) on the MuSiQue dataset and comprehensively evaluate their performance across a diverse set of benchmarks, including: (1) MuSiQue, for assessing in-domain performance; (2) Selected QA tasks from LongBenchV1 (Bai et al., 2023), for measuring

generalization capabilities on both single and multi-document question answering; and (3) Selected QA tasks from LongBenchV2 (Bai et al., 2024b), which feature longer contexts and more diverse domains. Compared to the baseline method of outcome supervision, our proposed process-supervised framework demonstrates superior performance on both the in-domain MuSiQue dataset (+13.6,+3.8 and +7.5 points for LLaMA/Qwen) and other QA tasks (+9.3, +8.1 and +10.4 points on average). These results validate the effectiveness of our approach across long-context scenarios spanning various domains and lengths.

Our contributions can be summarized as follows:

- We conduct, for the first time, a systematic examination to consolidate the effectiveness of CoT across most long-context scenarios of varying lengths and domains.

- We propose a long-context process-supervised framework comprising CoT sampling and a quality assessment protocol that efficiently ensures both answer correctness and process reliability of reasoning paths for long-context scenarios.

- Comprehensive experiments validate the superiority of our framework in both in-domain and generalization performance, and demonstrate the effectiveness of our quality assessment protocol.

2 Related Work

Long-Context Language Modeling Context length is one of the most fundamental properties of language models, determining the total amount of information they can process at once. Long-context language modeling seeks to extend this capacity, pushing the boundaries of models' context windows. Research for context window extension primarily follows two directions. The first is data-driven strategies, which focus on constructing training data that exhibits long-range dependency patterns (Gao et al., 2024; Fu et al., 2024; Chen et al., 2024a; Dubey et al., 2024; Xiong et al., 2024; Si et al., 2024; Wang et al., 2024a; Chen et al., 2024b; Bai et al., 2024a; Wu et al., 2024). The second is architecture-driven approaches, which enhance models' potential for processing long contexts by modifying components such as position encodings (Chen et al., 2023; Zhu et al., 2023; Peng et al., 2024; Ding et al., 2024) and attention mechanisms (An et al., 2024; Jin et al., 2024). Building upon these advances, current LCLMs have

---

[1] In this work, we use the terms *CoT* and *reasoning path* interchangeably. They both refer to the models' reasoning thoughts including the final answer.

demonstrated remarkable performance on simple long-context tasks, such as needle-in-haystack retrieval (Kamradt, 2023; Hsieh et al., 2024; Zhu et al., 2024). However, these models still struggle with more complex tasks, particularly long-context reasoning (Kuratov et al., 2024; Li et al., 2024b; Levy et al., 2024). Motivated by this challenge, our work investigates how to improve models' long-context reasoning capabilities by guiding them to generate high-quality Chains of Thought (CoT).

Process Supervision   Process supervision is a concept proposed to distinguish from outcome supervision (Uesato et al., 2022). In process supervision, models are not only guided toward correct answers but also taught appropriate reasoning patterns to reach these conclusions. A significant application of process supervision is in reinforcement learning (Uesato et al., 2022; Lightman et al., 2023; Luo et al., 2024), where trained reward models for process evaluation, known as Process Reward Models (PRMs), are used to supervise the execution process of policy models. Numerous studies have demonstrated that process supervision outperforms outcome supervision in domains requiring multi-step reasoning, such as mathematics (Wang et al., 2024b; Lightman et al., 2023). Beyond the reinforcement learning paradigm, many works directly utilize high-quality CoT, obtained either through model's self-rejection sampling (Zelikman et al., 2022; Singh et al., 2023; Zhang et al., 2024; Hosseini et al., 2024; Pang et al., 2024; Wang et al., 2024c) or sampling from other models, to train models' reasoning process via supervised fine-tuning (QwQ, 2024; Guo et al., 2025; Team et al., 2025). Our approach falls into the latter category. Most closely related to our work is SEALong (Li et al., 2024a). Our method differs from it in two significant aspects: (1) While SEALong focuses on unsupervised scenarios with limited improvements, we achieve substantial in-domain and generalization gains by building upon outcome supervision in supervised settings. (2) We specifically design a CoT quality assessment protocol tailored for long-context scenarios to ensure both accuracy and reliability of generated CoTs, whereas SEALong primarily relies on outcomes to select CoTs.

3   Efficacy of CoT in Long Context Tasks

Tasks   To comprehensively evaluate the effectiveness of CoT in long context scenarios, we curate a diverse set of tasks encompassing both synthetic

Tasks	Real-World			Synthetic		Total
	SQA	MQA	LICL	S-NIAH	MNR3	
#Data	126	90	41	96	96	449
AvgLen	49k	48k	76k	57k	57k	55k
Source	LongBenchV2			RULER	BABILong	-

Table 1: Statistics of the curated evaluation dataset. *AvgLen* is short for *Average Length*, which is measured by number of tokens according to the tokenizer of LLaMA3.1 (Dubey et al., 2024).

and real-world scenarios, covering a wide length range from 10k to 128k tokens. For synthetic scenarios, we select the single needle retrieval task (S-NIAH) from RULER (Hsieh et al., 2024) and the multi-needle reasoning task with 3 supporting facts (MNR3) from BABILong (Kuratov et al., 2024), which specifically focus on assessing models' retrieval and reasoning capabilities within extended contexts. Since synthetic datasets allow flexible length control, we sample an equal number of examples at lengths of {16,32,64,128}k for testing. For real-world scenarios, we select three representative tasks from LongBench-V2 (Bai et al., 2024b): Single-Document Question-Answering (SQA), Multi-Document Question-Answering (MQA), and Long In-Context Learning (LICL), covering various domains including Academic, Literary, Legal, Finance, etc. For each task type, we retain test samples with lengths less than 128k. Basic statistics of the curated evaluation dataset is presented in Table 1. In this section, we conduct a systematic investigation into the effectiveness of Chain-of-Thought in zero-shot long context scenarios. Below, we will first introduce our evaluation setting (§ 3.1), and then present the detailed experimental results (§ 3.2).

3.1   Evaluation Setup

Metrics   To facilitate evaluation, we structure all tasks into a multiple-choice QA format to standardize the model's output format, which enables us to directly use choice accuracy as the evaluation metric. It is worth noting that the original tasks in RULER and BABILong are not presented in a multiple-choice format. Therefore, we create multiple-choice QA pairs by combining the current sample's ground truth with ground truth answers from other samples to form the candidate options.

Models   To comprehensively analyze the effectiveness of CoT in long context scenarios, we conduct experiments using both open-
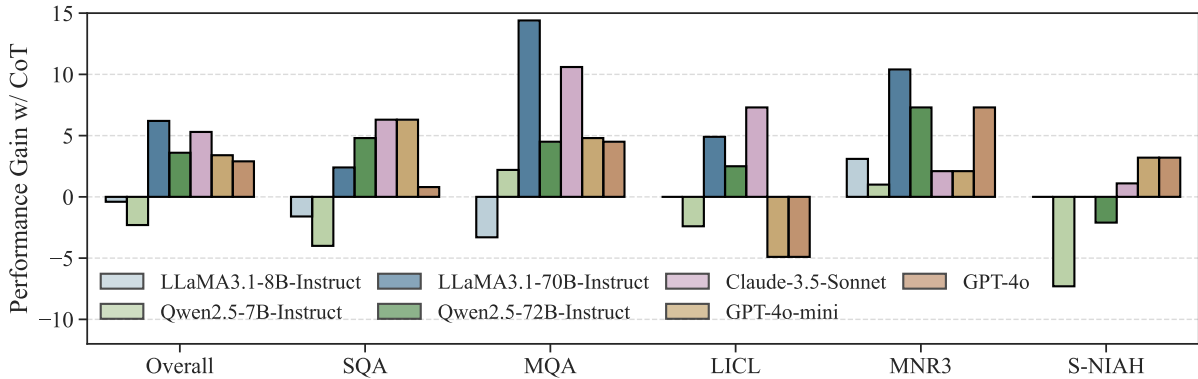
Figure 2: Performance gain of CoT for synthetic (MNR3, S-NIAH) and real-world (SQA, MQA, LICL) long context scenarios across all models. It is demonstrated that CoT particularly benefits proprietary and large-scale open-source models, and its effectiveness ranges across most long context scenarios, except for extremely easy retrieval tasks.

source and proprietary models of different scales and architectures, all supporting 128k context length. The tested models include LLaMA3.1-{8B,70B}-Instruct (Dubey et al., 2024), Qwen2.5-{7B,72B}-Instruct (Yang et al., 2024b), Claude-3.5-Sonnet (Claude, 2024), GPT-4o (GPT4o, 2024) and GPT-4o-mini (GPT4o, 2024). For ensuring fair benchmarking, the prompts employed in the evaluation strictly follow the two-step reasoning framework established in LongBench v2 (Bai et al., 2024b), as shown in Appendix A.

3.2 Observations

Figure 2 presents inference-time performance gain of CoT for both synthetic and real-world long context scenarios across all models.

First, we observe that **CoT particularly benefits proprietary and large-scale open-source models (Observation 1)**, as evidenced by the improvement on LLaMA3.1-70B-Instruct, Qwen2.5-70B-Instruct, Claude-3.5-Sonnet, GPT-4o-mini. By contrast, small scale models including LLaMA3.1-8B-Instruct and Qwen2.5-7B-Instruct fail to yield consistent performance gains from CoT. We hypothesize that this is due to smaller models' weaker overall capabilities, which may prevent them from generating high-quality reasoning paths to derive correct answers.

In addition, we find that **CoT is effective across most long context scenarios, except for extremely easy retrieval tasks (Observation 2)**. This is consolidated by the results that most models yield improvement on complex real-world scenarios (SQA, MQA, LICL) and synthetic scenarios that requires multi-hop reasoning over long context (MNR3), while showing no gain and even suffering

loss on the easiest single needle retrieval task (S-NIAH). We attribute this to the fact that S-NIAH is already sufficiently simple for most LCLMs, typically achieving 100% accuracy, and introducing CoT may actually create interference. Notably, while Sprague et al. (2024) indicates that CoT is only effective for short-context tasks involving symbolic operations and reasoning, our observations complement their findings and demonstrate the importance of CoT prompting for long-context tasks.

Further analysis explores how CoT benefits test samples of different lengths. We divide the test data into three length tiers: Short (<32k), Medium (32k-96k), and Long (>96k), with 199, 161, and 89 samples each. Table 2 presents the performance gains achieved with CoT across all large-scale models for each length interval. We observe that the Medium and Long groups generally show higher improvements with CoT compared to the Short group. This indicates that **CoT provides more benefits for longer tasks (Observation 3)**, possibly due to the increased need for implicit reasoning steps for information retrieval in longer contexts. We additionally provide a task-specific analysis across context lengths detailed in Appendix C.

4 Improving LCLMs via Self-Sampled Reasoning Paths

Section 3 consolidates the substantial benefits of CoT in long-context scenarios. Encouraged by these findings, we further explore how to enhance models' ability to generate high-quality CoTs in long-context scenarios, thereby improving their reasoning capabilities over long contexts. This section details the proposed process-supervised training framework, which comprises a self-sampling phase
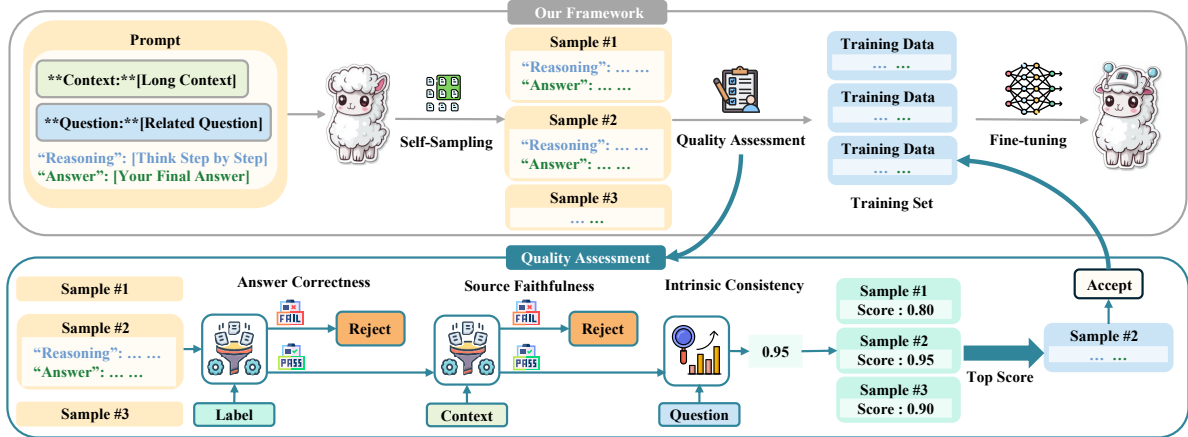
Figure 3: Our process-supervised framework for long-context tasks. We begin by sampling a diverse collection of N reasoning paths from the model. A quality assessment procedure consisting of three criteria is then applied to these samples to select high-quality training samples, which are then used for supervised fine-tuning.

Model	Short	Medium	Long
LLaMA3.1-70B-Instruct	+2.0	+9.9	+9.0
Qwen2.5-72B-Instruct	+1.0	+5.0	+6.7
Claude-3.5-Sonnet	+3.5	+9.3	+1.6
GPT-4o	-0.5	+5.5	+5.7
GPT-4o-mini	+4.0	+1.3	+6.2
Avg.	+2.0	+6.2	+5.8

Table 2: Performance gain of CoT on test samples of different length tiers: Short (<32k), Medium (32k-96k), and Long (>96k). CoT generally provides more benefits for longer samples (Medium & Long).

to collect reasoning paths, a quality assessment phase specifically designed for long-context scenarios to identify high-quality CoTs, and a training phase that performs supervised fine-tuning using the high-quality CoTs. The overall training process is illustrated in Figure 3.

4.1 Self-Sampling

We collect diverse reasoning paths via self-sampling, generating N variants per example to explore different reasoning strategies. During sampling, we employ a prompt (See Appendix A) that guides models through three key steps: (1) breaking down the question into manageable components, (2) identifying and citing relevant excerpts from the source text, and (3) deriving conclusions based on these excerpts. Crucially, we require models to explicitly cite source text by marking excerpts with '[Excerpt xxx]', where each excerpt must exactly match a portion of the original document. This strict matching requirement enables efficient

and reliable assessment of source faithfulness in the subsequent quality evaluation phase (Section 4.2).

4.2 CoT Quality Assessment

To identify high-quality CoT, we assess CoT quality via *answer correctness* and *process reliability*, ensuring alignment with ground truth and logical coherence.

Answer Correctness (AC) First, we require CoT to arrive at correct answers, filtering out failed reasoning paths. Specifically, we compute the F1-score between the CoT-derived final answer and the ground truth, retaining only reasoning paths with F1-scores above a threshold δ. This straightforward criterion ensures the basic effectiveness of the reasoning process.

Process Reliability Beyond answer correctness, we further require the reasoning process itself to be reliable. Evaluating the reliability of reasoning processes is particularly challenging in long-context scenarios, as it requires referencing extensive input text. Even when using LLMs capable of handling long inputs, ensuring assessment accuracy remains difficult and computationally expensive. To address this challenge, we decompose process reliability into two components: *source faithfulness* and *intrinsic consistency*. The former ensures reasoning faithfulness to the source text and can be efficiently measured through simple string matching with our novel design, while the latter focuses on CoT's inherent quality and can be assessed by LLMs without requiring additional context. This decomposition enables efficient evaluation of reasoning processes in long-context scenarios.

- **Source Faithfulness** Source Faithfulness requires reasoning paths to be grounded in the actual content present in the long context. This is crucial to prevent hallucination or the inclusion of irrelevant content that could contaminate the training data. By requiring models to provide relevant excerpts during sampling (Section 4.1), we can directly employ substring exact matching to verify citation faithfulness against the source text. This enables efficient assessment of CoT's information fidelity, allowing us to filter out reasoning paths that contain content inconsistent with the original text.

- **Intrinsic Consistency** After ensuring source faithfulness to the input text, we further examine the quality of CoT itself, which we term as Intrinsic Consistency. Specifically, a high-quality CoT should demonstrate logical coherence (appropriate question breakdown, logical use of information, and sound reasoning chain), completeness (primary reliance on retrieved information rather than model's internal knowledge), and conciseness (avoiding irrelevant or excessive details). Given the complexity of evaluating these dimensions, we employ LLM scoring with prompts detailed in Appendix A.

Finally, for each retained example, we select the CoT with the highest intrinsic consistency score for inclusion in the training data. In this way, we construct a high-quality self-sampled dataset for further model fine-tuning.

4.3 Supervised Fine-tuning

After collecting high-quality self-sampled reasoning paths, high-quality CoTs are used for supervised fine-tuning to improve long-context task performance. Specifically, we minimize the expected negative log-likelihood:

$$\mathcal{L}_\theta = \mathbb{E}_{(x,y)\sim\mathcal{D}}[-\log \mathcal{M}_\theta(y|x)]$$

where $\mathcal{D}$ denotes the training dataset, and $\mathcal{M}_\theta(y|x)$ denotes the conditional likelihood of CoT y under the model parameterized by θ.

5 Experiments

5.1 Experimental Setups

Training Data We conduct our self-training experiments on MuSiQue (Trivedi et al., 2022), a challenging multi-hop QA task that requires models to integrate information from multiple Wikipedia documents to answer questions. To better align with modern long-context scenarios, we extend the dataset's context length to 10-16k tokens following SEALong's method (Li et al., 2024a).

Candidate Models We select LLaMA-3.1-8B-Base (Dubey et al., 2024) and Qwen-2.5-{7B, 32B}-Base (Yang et al., 2024a) as the candidate models for CoT self-training. Notably, we opt for base versions rather than instruction-tuned variants, as the latter typically undergo sophisticated post-training procedures that may include CoT generation optimizations, which could compromise the reliability of our experimental results. To equip the base models with basic instruction-following capabilities, we perform a simple warmup process at first (details provided in *Implementation Details*). We also acknowledge the necessity of conducting experiments across models of varying scales. Our current computational resources remain insufficient to perform experiments on larger-scale models such as Llama-70B even considering LoRA fine-tuning (Hu et al., 2022).

Evaluation Setup To comprehensively analyze the effectiveness of our proposed self-training framework, we evaluate the trained models across a diverse set of datasets, including: (1) MuSiQue, for assessing in-domain performance; (2) Selected QA tasks from LongBenchV1 (Bai et al., 2023), including Qasper (Dasigi et al., 2021), HotpotQA (Yang et al., 2018), MultiFieldQA-En (Bai et al., 2023), and 2WikiMultihopQA (Ho et al., 2020), for measuring generalization capabilities on both single and multi-document question answering; and (3) Selected QA tasks from LongBenchV2 (Bai et al., 2024b), which feature longer contexts and more diverse domains. We employ F1-score and multiple-choice accuracy as evaluation metrics, depending on the format in which each task is structured. We also include the results from for LLaMA3.1-8B-Instruct, Qwen-2.5-{7B, 32B}-Instruct, GPT-4o, GPT-4o-mini for reference.

Implementation Details Our training dataset for MuSiQue comprises 3,300 examples in total. For the warmup stage, we first select 300 examples and obtain their reasoning paths from the instruction-tuned versions of candidate models, then fine-tune their corresponding base models on these CoT examples for 20 steps. In the self-training stage, we use the warmed-up models to generate reasoning paths for the remaining 3,000 examples. Specifically, we sample 30 CoTs per example and apply

Model	MuSiQue	QAs-LBV1				QAs-LBV2		Avg.
		HPQA	MFQA	Qasper	2WMQA	SQA	MQA	
LLaMA-3.1-8B-Instruct	45.3	57.3	53.8	42.9	64.0	29.3	32.2	46.4
LLaMA-3.1-8B-Base	12.6	21.2	29.9	13.4	19.9	1.2	1.7	14.3
w/ Outcome Supervision	47.0	50.0	44.4	32.1	37.1	17.1	14.8	34.7
w/ Ours	60.6(+13.6)	57.9(+7.9)	53.8(+9.4)	36.0(+3.9)	50.5(+13.4)	28.1(+11.0)	31.3(+16.5)	44.0(+9.3)
Qwen-2.5-7B-Instruct	39.4	57.5	48.7	43.0	54.2	34.2	33.0	44.3
Qwen-2.5-7B-Base	23.8	43.8	46.2	29.9	28.2	30.5	32.2	33.5
w/ Outcome Supervision	49.2	58.1	43.2	29.7	41.2	20.7	17.4	37.1
w/ Ours	53.0(+3.8)	57.0(-1.1)	45.6(+2.4)	38.4(+8.7)	58.1(+16.9)	30.5(+9.8)	33.9(+16.5)	45.2(+8.1)
Qwen-2.5-32B-Instruct	44.1	61.2	53.6	45.1	75.1	44.4	36.6	51.4
Qwen-2.5-32B-Base	28.8	46.4	41.7	27.7	43.2	41.7	34.8	38.0
w/ Outcome Supervision	43.4	57.6	36.5	23.7	54.9	33.7	29.5	40.1
w/ Ours(LoRA)	50.9(+7.5)	63.6(+6.0)	49.4(+12.9)	41.2(+13.5)	68.3(+13.4)	42.1(+8.4)	36.0(+6.5)	50.5(+10.4)
GPT-4o-mini	46.3	56.1	50.2	38.7	64.0	34.2	34.2	46.2
GPT-4o	55.8	65.8	54.8	45.4	74.8	46.0	47.2	55.7

Table 3: Performance comparison of different models on the selected QA tasks from LongBenchV1 (*QAs-LBV1*) and the longer and more domain-diverse QAs from LongBenchV2 (*QAs-LBV2*). *HPQA*, *MFQA*, *2WMQA*, *SQA*, *MQA* is short for HotpotQA, MultiFieldQA, 2WikiMultihopQA, Single-Document QA, Multi-Document QA, respectively. Numbers in parentheses show changes from outcome supervision baseline, (improvements in green, degradations in red). Our method not only excels on in-domain tasks (MuSiQue) but also exhibits strong generalization capabilities.

our proposed filtering strategy (Section 4.2) to construct the training set. The remaining examples are used to fine-tune the warmed-up models for 2 epochs. For the baseline model using outcome supervision, we directly fine-tune the warmed-up models on the complete 3000 examples for 2 epochs. Due to computational constraints, we employ LoRA (Hu et al., 2022) for 32B models. For fair comparison, we report the best performance achieved across the two epochs. Please refer to the Appendix D for more detailed hyperparameter configurations.

5.2 Main Results

Table 3 presents the performance of our models across the evaluated datasets.

CoT Supervision Beats Outcome Supervision First, we observe that our proposed method significantly outperforms the baseline method of outcome supervision on MuSiQue for *LLaMA-3.1-8B-Base* and *Qwen-2.5-{7B,32B}-Base*. The improvement is particularly pronounced for the *LLaMA-3.1-8B-Base* model, demonstrating a substantial performance gain of +13.6 points. Similarly, in the model's majority voting performance evaluation, our method also achieves a significant improvement (See Appendix B). This validates the overall effectiveness of CoT supervision, which not only guides models toward correct answers but also teaches them appropriate reasoning patterns to

Assess Criteria	MuSiQue	QAs-LBV1	QAs-LBV2	Avg.
LLaMA-3.1-8B				
AC check only	54.0	46.7	22.9	40.9
+ SF	54.8	46.3	25.1	41.4
+ SF + IC	**60.6**	**47.0**	**29.7**	**44.0**
Qwen-2.5-7B				
AC check only	49.8	49.7	29.2	43.9
+ SF	50.9	**50.1**	31.8	45.0
+ SF + IC	**53.0**	49.8	**32.2**	**45.2**

Table 4: Ablation experiments on our proposed CoT quality assessment protocol. *AC*, *SF*, and *IC* are short for *Answer Correctness*, *Source Faithfulness*, and *Intrinsic Consistency*, respectively. Other notations follow Table 3. The results show that each assessment criterion contributes positively to the model's performance.

reach these conclusions.

Superior Generalizability Furthermore, Table 3 compares the generalization performance of our method against the baseline across various tasks. Results show that our method consistently achieves superior performance, both on other multi-document QA tasks from LongBenchV1 and on tasks from LongBenchV2 featuring longer contexts and more diverse domains. Specifically, we observe average performance gains of +9.3, +8.1 and +10.4 points for LLaMA and Qwen models respectively, reaching performance levels comparable to GPT-4o-mini. More importantly, the performance improvement observed on Qwen-2.5-32B suggests
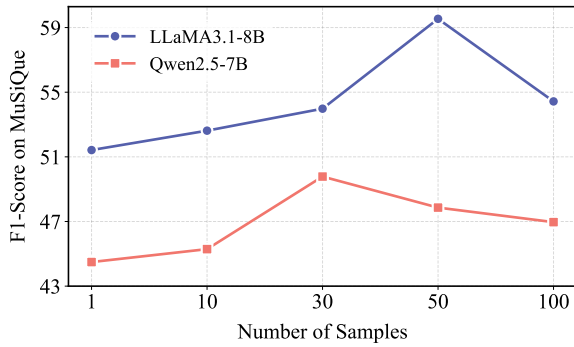
Figure 4: Impact of sampling size on model performance on MuSiQue.

Figure 5: The impact of CoTs of different quality on model performance. Notations align with Table 3.

that our process supervision method remains effective with increasing model size without showing diminishing returns. These results demonstrate that our method exhibits strong generalization capabilities on out-of-domain tasks, bringing improvements across long-context scenarios of varying formats, domains, and context lengths.

5.3 Ablation Study

Efficacy of Our Quality Assessment Protocol We further examine the effectiveness of our assessment protocol in selecting high-quality CoTs as training data. Since maintaining answer correctness is a fundamental requirement and de facto standard in common practice, its effectiveness is self-evident. We thus focus on examining the criteria for process reliability: source faithfulness and intrinsic consistency. As shown in Table 4, incorporating each assessment criterion brings positive benefits, with the highest performance achieved when all three criteria work together. For the LLaMA-3.1-8B model, compared to using answer consistency check alone, adding source faithfulness check leads to an average gain of +0.5 points, while further incorporating intrinsic consistency scoring brings an additional +2.6 points. Similarly, for the Qwen-2.5-7B model, applying all three assessment criteria improves the average score by +1.3 points. This consistent improvement across both models underscores the necessity of each component in ensuring the quality of self-sampled reasoning paths.

6 Analysis

Impact of Sampling Size First, we analyze how the number of sampled candidate CoTs per training example affects model performance. As expected, a larger sampling size provides more diverse candidates, potentially enabling the selection of higher-quality CoTs. Figure 4 illustrates the impact of
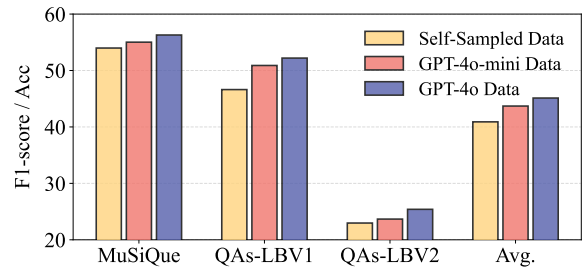
sampling size on model performance on the test set of MuSiQue. Interestingly, we observe that model performance first increases and then decreases with the increasing sampling size. Performance peaks at 30 samples for Qwen-2.5-7B and 50 for Llama-3.1-8B. We hypothesize that this phenomenon might be due to the increased difficulty in maintaining consistent quality assessment with larger candidate pools.

Impact of CoT Source Models Beyond self-sampling CoTs from the base model, we investigate the effectiveness of training with CoTs directly sampled from more capable models (GPT-4o-mini and GPT-4o), as shown in Figure 5. We observe that GPT-4o-generated CoTs perform best across scenarios: the in-domain MuSiQue task, other QA tasks from LongBench V1 and V2, followed by GPT-4o-mini and then self-sampled CoTs. Combined with our ablation results in Table 4, these findings demonstrate that improving CoT quality, whether through careful selection from self-sampled candidates or direct sampling from stronger models, effectively enhances model capabilities in long-context scenarios.

7 Conclusion

In this study, we propose a process-supervised framework to enhance the long-context reasoning capabilities of LLM by guiding them to generate high-quality reasoning paths. By conducting experiments on various long-context benchmarks, we demonstrate that our framework significantly improves model performance on both in-domain and out-domain tasks, highlighting the generalization capabilities of our method. Our experiments underscore the critical role of process-supervised reasoning paths in enhancing long-context language models, enabling them to navigate and reason over extensive textual information with greater accuracy and reliability. By reducing the reliance on human annotations and leveraging self-sampled data, our

approach offers a scalable solution for enhancing LLMs' reasoning abilities.

Limitations

Our work is still limited in some aspects. Although we have conducted experiments on Llama-3.1-8B and Qwen-2.5-{7B,32B} to verify the validity of our proposed framework, we lack experiments on larger models (e.g., 70B and 140B parameters) due to limited computational resources. Similarly, due to GPU memory constraints, we can only fine-tune our model on sequences up to 16K tokens in length. We believe that fine-tuning on longer sequences would lead to better performance on long-text tasks. In future work, we plan to address these limitations and further explore the effectiveness of our framework on larger models and longer sequences.

Ethics Statement

This work fully complies with the ACL Ethics Policy. We declare that there are no ethical issues in this paper, to the best of our knowledge.

Acknowledgement

References

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. 2024. Why does the effective context length of llms fall short? *arXiv preprint arXiv:2410.18745*.

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

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Egor Bogomolov, Aleksandra Eliseeva, Timur Galimzyanov, Evgeniy Glukhov, Anton Shapkin, Maria Tigina, Yaroslav Golubev, Alexander Kovrigin, Arie van Deursen, Maliheh Izadi, et al. 2024. Long code arena: a set of benchmarks for long-context code models. *arXiv preprint arXiv:2406.11612*.

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

Longze Chen, Ziqiang Liu, Wanwei He, Yinhe Zheng, Hao Sun, Yunshui Li, Run Luo, and Min Yang. 2024a. Long context is not long at all: A prospector of long-dependency data for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8222–8234, Bangkok, Thailand. Association for Computational Linguistics.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *ArXiv*, abs/2306.15595.

Zhi Chen, Qiguang Chen, Libo Qin, Qipeng Guo, Haijun Lv, Yicheng Zou, Wanxiang Che, Hang Yan, Kai Chen, and Dahua Lin. 2024b. What are the essential factors in crafting effective long context multi-hop instruction datasets? insights and best practices. *ArXiv*, abs/2409.01893.

Team Claude. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. In *Forty-first International Conference on Machine Learning*.

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

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. In *Forty-first International Conference on Machine Learning*.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2024. How to train long-context

language models (effectively). *arXiv preprint arXiv:2410.02660*.

Team GPT4o. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

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

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-STar: Training verifiers for self-taught reasoners. In *First Conference on Language Modeling*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Selfextend llm context window without tuning. In *Forty-first International Conference on Machine Learning*.

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

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

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. 2024a. Large language models can self-improve in long-context reasoning. *arXiv preprint arXiv:2411.08147*.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024b. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. Nolima: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason E Weston. 2024. Iterative reasoning preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Team QwQ. 2024. Qwq. https://qwenlm.github.io/blog/qwq-32b-preview/.

Shuzheng Si, Haozhe Zhao, Gang Chen, Yunshui Li, Kangyang Luo, Chuancheng Lv, Kaikai An, Fanchao Qi, Baobao Chang, and Maosong Sun. 2024. Selecting influential samples for long context alignment via homologous models' guidance and contextual awareness measurement.

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. 2023. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Liang Wang, Nan Yang, Xingxing Zhang, Xiaolong Huang, and Furu Wei. 2024a. Bootstrap your own context length. *arXiv preprint arXiv:2412.18860*.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.

Tianduo Wang, Shichen Li, and Wei Lu. 2024c. Self-training with direct preference optimization improves chain-of-thought reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11917–11928.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wenhao Wu, Yizhong Wang, Yao Fu, Xiang Yue, Dawei Zhu, and Sujian Li. 2024. Long context alignment with short instructions and synthesized positions. *arXiv preprint arXiv:2405.03939*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2024. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4643–4663, Mexico City, Mexico. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin,

Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024b. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STar: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

Dan Zhang, Sining Zhoubian, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. Rest-mcts*: Llm self-training via process reward guided tree search. *arXiv preprint arXiv:2406.03816*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. LongEmbed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2023. Pose: Efficient context window extension of llms via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*.

A Prompts

Prompt for CoT Analysis #1

Please read the following text and answer the questions below.
<text>
DOC
</text>
What is the correct answer to this question: Q
Choices:
(A) C_A
(B) C_B
(C) C_C
(D) C_D
Let's think step by step:

Prompt for CoT Analysis #2

Please read the following text and answer the questions below.
The text is too long and omitted here.
What is the correct answer to this question: Q
Choices: (A) C_A
(B) C_B
(C) C_C
(D) C_D
Let's think step by step: COT
Based on the above, what is the single, most likely answer choice? Format your response as follows: "The correct answer is (insert answer here)".

You are given a long document such as a story, meeting script, a news article, etc, and a question. Your task is to answer the question based on the information provided in the document. You should follow the instructions below to provide an accurate reasoning path, as well as a concise answer to the question:

Instructions:
Step 1. **Reasoning:** First retrieve all relevant information, then deduce the correct answer. Begin by carefully reading the provided context. Identify and extract all relevant information that is directly related to the question. Be succinct and only extract the most important excerpts that will help you answer the question. Finally, deduce the correct answer based on the retrieved information.
Step 2. **Answer:** Using the information you have retrieved, and your deduction, answer the question as concisely as you can, using a single phrase or sentence if possible. Ensure that your answer should be brief and to the point.
Step 3. **Format Your Response:** Present your response in JSON format, comprising two components: "reasoning" and "answer". The "reasoning" section should detail your thought process, including the breakdown of the question, the relevant excerpts (indicated by [Excerpt xxx] at the start), and the derived conclusion. Ensure that each excerpt is an exact match to the original document. Limit the number of excerpts to a maximum of 10. The "answer" part should contain your final answer to the question, as concise and to the point as possible.

Illustrative Examples:
Example #1:
Context: [... Saltram is living with the Mulvilles at Wimbledon ... He is not working or producing anything ... He is idle and dependent on others ...]
Question: What is Saltram's living situation?
Response:
{{
"reasoning": "Let me first retrieve relevant excerpts from the document, then answer the question. The question asks about Saltram's living situation. In the document, I can first locate that [Excerpt 1] 'Saltram is living with the Mulvilles at Wimbledon'. Additionally, it is mentioned that [Excerpt 2] 'He is not working or producing anything' and [Excerpt 3] 'He is idle and dependent on others'. From these excerpts, I can deduce that Saltram is a guest in the home of the Mulvilles.",
"answer": "He is a guest in the home of the Mulvilles."
}}

Example #2:
Context: [... The Collegian is the bi-weekly official student publication of Houston Baptist University in Houston, Texas ... Houston Baptist University, affiliated with the Baptist General Convention of Texas, offers bachelor's and graduate degrees. It was founded in 1960 ...]
Question: When was the institute that owned The Collegian founded?
Response:
{{
"reasoning": "Let me first retrieve relevant excerpts from the document, then answer the question. The question asks about the founding date of the institute that owned The Collegian. In the document, I can first locate that [Excerpt 1] 'The Collegian is the bi-weekly official student publication of Houston Baptist University in Houston, Texas', so I need to look for information about Houston Baptist University. I find that [Excerpt 2] 'Houston Baptist University was founded in 1960'. Therefore, the institute that owned The Collegian was founded in 1960.",
"answer": "1960"
}}

Now, based on the context provided below, answer the question as concisely as you can, using a single phrase or sentence if possible.
Context: {context}
Question: {question}
Response:

Prompt for LLM Scoring

[Question]
{question}

[The Start of Assistant's Reasoning Path]
{reasoning}
[The End of Assistant's Reasoning Path]

[System]
We would like to request your feedback on the quality of the reasoning process in the given response. The model receives a long text input and a complex question. Its task is to retrieve relevant information from the long text (marked as [Excerpt xxx] and enclosed in ") based on the question's requirements and provide the correct answer. Above, we have provided both the question and the model's reasoning process. While the model's final answer is correct, we need you to evaluate whether its reasoning process is sound.

Please assess the model's reasoning process based on the following aspects:

1. Logical Coherence:
- The model should break down the question appropriately
- The use of retrieved information should follow logical patterns
- The chain of reasoning from retrieved information to the final answer should be sound

2. Completeness:
- The reasoning process should primarily rely on information retrieved from the text ([Excerpts xxx])
- The model should not heavily depend on its own knowledge base

3. Conciseness:
- Only information relevant to answering the question should be retrieved
- The model should avoid listing excessive or irrelevant information

Please rate whether this reasoning path is suitable for the question. The assistant receives an overall score on a scale of 1 to 100, where a higher score indicates better overall performance.
Please note that if the assistant's reasoning process fully meets the above criteria, its overall rating should be full marks (100).
Please first provide a comprehensive explanation of your evaluation, avoiding any potential bias.
Then, output a line indicating the score of the Assistant.

PLEASE OUTPUT WITH THE FOLLOWING FORMAT, WHERE THE SCORE IS ON A SCALE OF 1 TO 100 BY STRICTLY FOLLOWING THIS FORMAT: "[[score]]", FOR EXAMPLE "Rating: [[100]]":
<start output>
Evaluation evidence: your evaluation explanation here, no more than 100 words
Rating: [[score]]
<end output>

Now, start your evaluation:
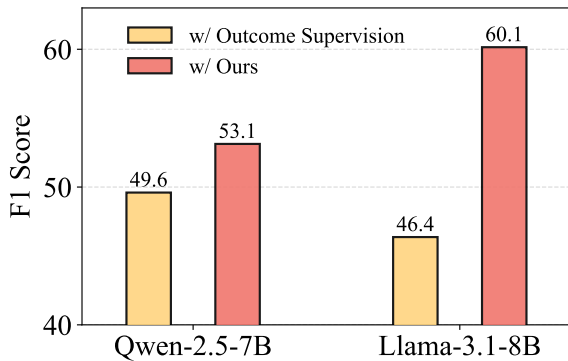
B Majority Voting Performance

Figure 6: Majority voting results on MuSiQue, w/ outcome supervision and w/ our proposed method. The models align with Table 3.

The majority voting results of Llama-3.1-8B and Qwen-2.5-7B are shown in Figure 6. We can see that for both models, our method shows a significant improvement compared to outcome supervision in the majority voting setting. This is consistent with the phenomena we observed in Table 3.

C Task-Specific Context Length Analysis

Model	Short	Medium	Long
LLaMA3.1-70B-Instruct	+1.3	+2.8	+4.0
Qwen2.5-72B-Instruct	+4.3	+5.5	+4.5
Claude-3.5-Sonnet	+3.2	+10.8	+2.8
GPT-4o	+0.2	+0.6	+2.6
GPT-4o-mini	+6.5	+3.3	+13.2
Avg.	+3.1	+4.6	+5.4

Table 5: Performance gain of CoT on SQA task of different length tiers: Short (<32k), Medium (32k-96k), and Long (>96k).

To address potential interactions between task types and context lengths in our CoT evaluation presented in Section 3, we provide supplementary results for the SQA task across context length tiers in Table 5. These results align with our core observations in Section 3.2, showing that CoT generally provides greater benefits for longer contexts even when examining a single task type. The pattern holds despite some model-specific variations, with the average improvement increasing from +3.1 (Short) to +5.4 (Long).

D Additional Implementation Details

The following provides detailed hyperparameter specifications for both the warmup stage and self-training stage:

Warmup Stage $lr = 1e^{-5}$, batch_size = 32
Filtering Framework Sampling Temperature = 0.7, Answer Consistency threshold = 1.0, CoT quality assessed via GPT-4o-mini
Self-Training Stage $lr = 5e^{-6}$, batch_size = 32
LoRA Configuration Rank = 128, α = 128, dropout = 0.05 (applied to all attention and feed-forward layers)