LONGFAITH: Enhancing Long-Context Reasoning in LLMs with Faithful Synthetic Data

Cehao Yang[1,2*], **Xueyuan Lin**[1,2,4*], **Chengjin Xu**[1,3*], **Xuhui Jiang**[1,3],
Shengjie Ma[1], **Aofan Liu**[1], **Hui Xiong**[2†], **Jian Guo**[1†]
[1]IDEA Research, International Digital Economy Academy
[2]Hong Kong University of Science and Technology (Guangzhou)
[3]DataArc Tech Ltd.
[4]Hithink RoyalFlush Information Network Co., Ltd
{yangcehao,linxueyuan,xuchengjin,jiangxuhui,mashengjie,liuaofan,guojian}@idea.edu.cn
{cyang289,xlin058}@connect.hkust-gz.edu.cn, xionghui@ust.hk

Abstract

Despite the growing development of long-context large language models (LLMs), data-centric approaches relying on synthetic data have been hindered by issues related to faithfulness, which limit their effectiveness in enhancing model performance on tasks such as long-context reasoning and question answering (QA). These challenges are often exacerbated by misinformation caused by lack of verification, reasoning without attribution, and potential knowledge conflicts. We propose LONGFAITH, a novel pipeline for synthesizing faithful long-context reasoning instruction datasets. By integrating ground truth and citation-based reasoning prompts, we eliminate distractions and improve the accuracy of reasoning chains, thus mitigating the need for costly verification processes. We open-source two synthesized datasets—LONGFAITH-SFT and LONGFAITH-PO—which systematically address multiple dimensions of faithfulness, including verified reasoning, attribution, and contextual grounding. Extensive experiments on multi-hop reasoning datasets and LongBench demonstrate that models fine-tuned on these datasets significantly improve performance. Our ablation studies highlight the scalability and adaptability of the LONGFAITH pipeline, showcasing its broad applicability in developing long-context LLMs.

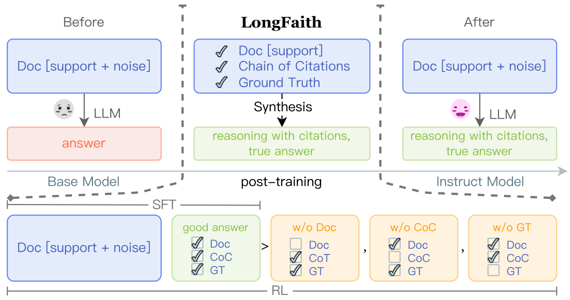

Figure 1: A brief introduction of LONGFAITH. Synthesized long-context reasoning instruction sets and preference datasets are fed into the post-training stage of downstream LLMs.

1 Introduction

Long-context processing ability has emerged as a significant challenge for large language models (LLMs) (Shi et al., 2023; Liu et al., 2024; Wu et al., 2024; Levy et al., 2024), especially arises when models process extensive textual information, making it hard to recognize relevant evidence and address downstream tasks such as question answering (QA), summarization, and complex reasoning (Bai et al., 2023, 2024b; Zhang et al., 2024d; Hsieh et al., 2024; Yen et al., 2024). A variety of model-centric methods have been proposed to extend the length of context windows in LLMs (Chen et al., 2023a,b; Peng et al., 2023; Han et al., 2024; Ding et al., 2024). Additionally, many data-centric methods, such as synthesizing long-context understanding instructions for fine-tuning, have gained attention for enhancing models' ability to handle and utilize extended contexts (Xiong et al., 2023; An et al., 2024; Fu et al., 2024; Bai et al., 2024a; Li et al., 2024b; Gao et al., 2024; Chen et al., 2024; Zhang et al., 2024c; Li et al., 2024a; Jiang et al., 2025).

Despite the improvements in downstream QA performance enabled by synthetic long-context reasoning instructions, concerns remain regarding the faithfulness of such generated data. Specifically: (1) **Misinformation due to lack of verification**: existing methods often generate QA pairs without rigorous rule-based verification. For instance, (Chen et al., 2024; Zhang et al., 2024c; Li et al., 2024a) directly synthesize QA pairs using LLMs while bypassing verification, whereas (Zhang et al., 2024c) relies on AI-generated feedback in soft dimensions rather than human annotation. (2) **Reasoning without attribution**: prompting LLMs to generate responses with citation, such as using *Chain-of-Citation (CoC)* prompting (Li et al.,

*Equal contribution.
†Corresponding authors.
‡Our code implementation and datasets can be accessed at https://github.com/IDEA-FinAI/LongFaith.

Findings of the Association for Computational Linguistics: ACL 2025, pages 3236–3256
July 27 - August 1, 2025

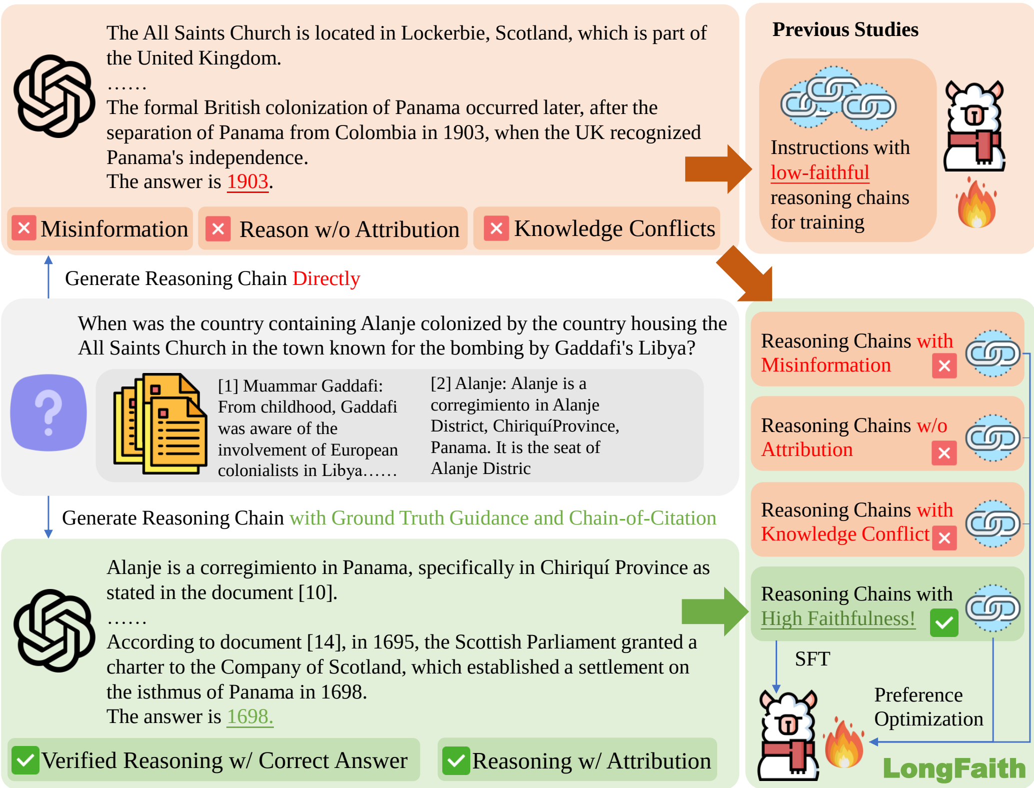

Figure 2: Overview of LongFaith pipeline for synthesizing faithful long-context reasoning instruction and preference datasets. Comparing generated reasoning chains with misinformation, lack of attribution, and knowledge conflicts, LongFaith generates ground truth guidance prompting by chain-of-citation to build LongFaith-SFT. Fine-grained faithfulness is modeled by optimization on our preference datasets LongFaith-PO.

2023, 2024b; Fierro et al., 2024; Huang et al., 2024; Berchansky et al., 2024; Gao et al., 2023) can enhance the credibility and interpretability of model outputs under long-context QA tasks (Gao et al., 2023; Zhang et al., 2024b), yet most prior works ignore to incorporate this technique during their synthesis of training instruction pairs. (3) **Potential knowledge conflicts**: some approaches (Bai et al., 2024a; Zhang et al., 2024c; Chen et al., 2024) over-rely on the *Self-Instruct* technique (Wang et al., 2022) to generate QA pairs, encouraging models to rely on parametric knowledge rather than grounding reasoning in explicit contextual evidence (Xu et al., 2024). Additionally, (Zhang et al., 2024c) feeds the query and response to a short-context reward model ignoring the long context to score, purely relying on the parametric knowledge inside LLM. These limitations underscore the necessity for a more robust pipeline that ensures the faithfulness of long-context instructions synthesis.

We propose **LongFaith**, a novel pipeline for synthesizing faithful long-context reasoning instruction datasets. We incorporate ground truth directly into the prompt for synthesizing long-context reasoning chains, which comprise supporting facts and the correct answer, and prompt LLMs to reason with attributions. This method ensures the faithfulness of synthesized reasoning chains without requiring costly verification by a curated rule-based evaluator, LLM-as-a-judge (Gu et al., 2024) or human annotator. We open-source **LongFaith-SFT**, synthesized under the guidance of ground truth and CoC prompting. We leverage the faithful long-context reasoning chains with attributions for training, leading to performance improvements after fine-tuning Llama-3.1-8B-Instruct. Additionally, we synthesize preference datasets by sampling preference pairs around fine-grained faithfulness: (1) encouraging model to reason with attributions; (2) encouraging model to learn on verified reasoning chains; and (3) encouraging model to reason with contextual documents grounded. We open-source **LongFaith-PO**, synthesized by various LLMs in different sizes, which integrates all three faithfulness dimensions for preference optimization. We leverage these faithful preferred instruction pairs for training Llama-3.1-8B-Instruct, achieving performance improvements on the multi-hop reasoning dataset and LongBench (Bai et al., 2023).

Our main contributions are as follows: (1) We introduce LONGFAITH, a novel pipeline for synthesizing faithful long-context reasoning instruction data. (2) We open-source LONGFAITH-SFT and LONGFAITH-PO, two synthesized datasets that are systematically designed considering multiple dimensions of faithfulness. (3) We conduct extensive experiments on two types of datasets (comprising eight sub-tasks) to show that models trained on LONGFAITH datasets can improve in long-context reasoning and QA tasks. (4) Our ablation studies illustrate the scaling potential and adaptability of LONGFAITH pipeline, underscoring its broad applicability in the development of long-context LLMs.

2 Related Work

Long-Context Utilization. Amounts of studies focus on enhancing LLMs to better utilize long-context information. Model-centric approaches, for instance, optimizations on attention mechanism aim to capture specific sequential features (Beltagy et al., 2020; Ding et al., 2023; Chen et al., 2023b; Han et al., 2024), while positional interpolation techniques are utilized to scale positional encoding while ensuring valid index ranges (Zhu et al., 2023; Chen et al., 2023a; Ding et al., 2024; Peng et al., 2023; Beltagy et al., 2020). In addition, data-driven approaches also gain popularity, emphasizing high-quality data synthesis for fine-tuning to improve LLMs' long-context processing capabilities. For example, (Xiong et al., 2023; Gao et al., 2024) employ long-sequence continuous pre-training on foundation models, while (Fu et al., 2024) explores the impact of pre-training data composition and balance. Additionally, works on SFT with synthetic instructions (An et al., 2024; Bai et al., 2024a; Li et al., 2024b; Chen et al., 2024) not only consider long-context understanding but also strengthen multi-hop reasoning capabilities. Lastly, preference optimization approaches (Zhang et al., 2024c; Li et al., 2024a) generate fine-grained pair-wise preference instruction sets and incorporate training techniques (Rafailov et al., 2024; Hong et al., 2024). From the perspective of improving the faithfulness of synthetic data, our work effectively addresses the shortcomings of prior studies in this area.

Faithful Reasoning. Hallucination in LLMs presents a major challenge in knowledge-intensive tasks (Zhang et al., 2023; Huang et al., 2023). Recent work has focused on enhancing faithful reasoning, where the goal is to trace the LLM's generated content back to reliable sources and ensure its factual grounding. (Berchansky et al., 2024; Li et al., 2023, 2024b) aim to improve the identification and verification of attributions by focusing on generating reasoning outputs that link claims to specific sources. Benchmarks such as (Gao et al., 2023; Yue et al., 2023) evaluate the quality of citations and highlight the limitations of current systems in providing citation support to ensure more reliable output. Additionally, integrating external knowledge sources has gained attention, which use retrieval-augmented generation (RAG) methods to facilitate deep and faithful reasoning (Sun et al., 2023; Ma et al., 2024). Our LONGFAITH is motivated by previous work, towards faithful reasoning in long-context reasoning tasks.

3 LongFaith

In this section, we present an exposition of LONGFAITH pipeline. Specifically, we explain how it synthesizes LONGFAITH-SFT for supervised fine-tuning and LONGFAITH-PO for preference optimization from the perspective of faithfulness.

Synthesize Reasoning Chains with High Faithfulness. Previous studies (Bai et al., 2024a; Chen et al., 2023b, 2024; Zhang et al., 2024c; Li et al., 2024a) tend to directly distill synthesized long-context QA and reasoning instructions for training without filtering out incorrect information. These low-faithfulness synthesized data limit the performance improvements of the trained models. In response to this challenge, LONGFAITH integrates ground truth into the synthesized reasoning chains. For a sample S from the training set $S = (Q, D, A, F)$, where Q is the reasoning question, D is the full document used for querying, A is the correct answer, and F represents the supporting facts where $F \in D$. We use the LLM M_{syn} to synthesize the reasoning chain as follows:

$$O_c = M_{\text{syn}}(P_{\text{coc}}, Q, F, A) \quad (1)$$

Here, O_c represents the output of M_{syn}, which is a step-by-step reasoning chain. The prompt P_{coc} utilizes a chain-of-citation (Li et al., 2024b) prompting approach, requiring the model to reason with attribution (e.g., "Let's reason step by step while citing the document using [1], [2], etc."). The prompt template is shown in Figure 7.

LONGFAITH-SFT Dataset. Towards training a downstream LLM to reason with high faithfulness for a long-context QA task, we construct the dataset for supervised fine-tuning, where each instruction pair is built as follow:

$$I_{\text{sft}} = \{\text{input} = (P_{\text{coc}}, Q, D), \text{output} = O_c\} \quad (2)$$

Synthesize Reasoning Chain with Questionable Faithfulness. To model fine-grained preferences, we address three challenges that affect the faithfulness of synthesized instructions: (1) misinformation due to lack of verification, (2) reasoning without attribution, and (3) potential knowledge conflicts. We synthesize reasoning chains with questionable faithfulness, including **reasoning chains with misinformation** as follows:

$$O_m = M_{\text{syn}}(P_{\text{coc}}, Q, D) \quad (3)$$

Since there is no ground truth to guide the synthesis, the output O_m may contain errors in reasoning, as illustrated in Figure 8, where the model generates an incorrect answer of "1903" instead of the correct answer "1698". This hallucination is common in synthesized data from previous works unless rules or human experts are involved in filtering (Li et al., 2024b). Next, we synthesize **reasoning chains without attribution**:

$$O_{\text{cot}} = M_{\text{syn}}(P_{\text{cot}}, Q, F, A) \quad (4)$$

Here, the CoT (Wei et al., 2022) prompting only requires the model to provide step-level reasoning, but as shown in Figure 9, reasoning without attribution not only loses interpretability and credibility (Gao et al., 2023; Li et al., 2023), but our results in Tab. 4 (Sec. 4) also demonstrate that CoT prompting performs worse than CoC. Finally, we synthesize **reasoning chains with potential knowledge conflicts**:

$$O_{\text{kc}} = M_{\text{syn}}(P_{\text{cot}}, Q, A) \quad (5)$$

Since no context is provided, the model relies solely on its parametric knowledge for reasoning, as shown in Figure 10, where the model states, "Panama was not colonized by the United Kingdom; Panama was colonized by Spain," based on internal parametric knowledge rather than the contextual documents. Previous studies (Zhang et al., 2024c) using short-context reward models observes performance degradation by ignoring long-context materials, highlighting the limitation of knowledge conflicts in affecting LLM's performance in long-context QA and reasoning tasks.

Synthesis of Reasoning Chains					
Models	**Prompt**	**w/ GT**	**w/ Doc**	**Output**	**Size**
Q-7B	CoC	✓	✓	1	1-8K
Q-7B	CoT	✓	✓	2	1-8K
Q-7B	CoC	✗	✓	3	1-8K
Q-7B	CoT	✓	✗	4	1-8K
L8,L70,G	CoC	✓	✓	5	2K
L8,L70,G	CoT	✓	✓	6	2K
L8,L70,G	CoC	✗	✓	7	2K
L8,L70,G	CoT	✓	✗	8	2K
Datasets for Supervised Fine-tuning					
Name	**Models**	**Instruction**	**Output**		**Size**
LF-SFT	*Q-7B*	CoC	1		1-8K
w/o CoC	*Q-7B*	CoT	2		1-8K
w/o GT	*Q-7B*	CoC	3		1-8K
w/o Doc	*Q-7B*	CoC	4		1-8K
LF-SFT	*L8,L70,G*	CoC	5		2K
w/o CoC	*L8,L70,G*	CoT	6		2K
w/o GT	*L8,L70,G*	CoC	7		2K
w/o Doc	*L8,L70,G*	CoC	8		2K
Datasets for Preference Optimization					
Name	**Models**	**Instruction**	**Chosen**	**Rejected**	**Size**
w/ CoC	*Q-7B*	CoC	1	2	1-8K
w/ GT	*Q-7B*	CoC	1	3	1-8K
w/ Doc	*Q-7B*	CoC	1	4	1-8K
LF-PO	*Q-7B*	CoC	1	2,3,4	1-8K
w/ CoC	*L8,L70,G*	CoC	5	6	2K
w/ GT	*L8,L70,G*	CoC	5	7	2K
w/ Doc	*L8,L70,G*	CoC	5	8	2K
LF-PO	*L8,L70,G*	CoC	5	6,7,8	2K

Table 1: Statistics of synthesized datasets for SFT and PO. We first synthesize large-scale reasoning chains and then refactor them to datasets, where the second stage does not require llm inference. ***Q-7B*** means *Qwen2.5-7B-Instruct*, ***L8*** means *Llama-3.1-8B-Instruct*, ***L70*** means *Llama-3.1-70B-Instruct* and ***G*** means *GPT-4o mini*. **GT** means Ground Truth, **CoC** means chain-of-citation, **Doc** means contextual documents, and **LF** means LONGFAITH. 1-8K includes {1K, 2K, 4K, 8K}.

LONGFAITH-PO Dataset. Towards training a downstream LLM to address three challenges above in long-context reasoning, we force the LLM to learn reasoning with high faithfulness while rejecting outputs of questionable faithfulness:

$$\begin{aligned} I_{\text{po}} = \{&\text{input} = (P_{\text{coc}}, Q, D), \\ &\text{chosen} = O_c, \text{rejected} = O_r\} \end{aligned} \quad (6)$$

where O_r is a combination of (O_m, O_{cot}, O_{kc}), or a subset of them.

Llama-3.1-8B-Instruct	MuSiQue		2Wiki		HotpotQA		Qasper(S)		MFQA-En(S)		MuSiQue(M)		2Wiki(M)		HotpotQA(M)	
	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM
Zero-Shot Prompting																
+ CoT	15.9	56.8	34.0	83.8	20.8	78.6	3.2	22.0	5.7	29.3	14.1	43.5	30.1	77.0	13.4	60.5
+ CoC	25.8	64.2	43.6	86.2	32.7	76.6	4.6	26.0	7.0	32.7	11.8	41.0	28.1	79.5	19.9	58.0
Superivised Fine-tuning																
+ LongAlpaca	21.6	50.2	47.8	80.4	32.7	76.6	5.7	25.0	5.8	30.7	8.5	48.5	25.4	77.0	12.5	61.0
+ LongAlign	24.8	48.4	55.6	84.2	51.0	79.2	6.5	24.0	10.7	**38.7**	15.0	40.0	33.4	76.5	35.8	61.0
+ MuSiQue-Attribute	13.9	19.2	23.9	49.6	20.2	37.2	10.0	11.5	8.3	12.0	15.2	26.5	21.2	55.0	25.6	41.0
+ LongMIT	4.9	33.0	3.3	58.0	10.1	63.6	9.5	18.5	5.6	30.0	7.5	29.0	3.6	55.5	23.7	50.0
+ LongReward-SFT	6.2	48.4	23.3	80.0	15.6	74.2	2.6	22.5	0.5	34.0	1.1	43.0	6.6	71.5	8.9	54.0
+ SeaLong-SFT	31.3	64.6	55.8	89.2	59.4	**83.0**	14.5	26.0	18.6	31.3	24.1	**59.5**	34.1	84.5	37.3	69.0
+ LongFaith-SFT	56.8	62.8	**73.8**	85.6	**70.5**	80.8	36.9	29.5	**47.0**	32.0	50.1	56.5	63.9	82.0	53.1	68.0
Preference Optimization																
+ LongReward-PO	3.3	46.0	14.3	76.6	8.9	71.2	1.6	21.0	0.1	32.7	0.0	37.5	4.4	67.0	3.3	53.0
+ SeaLong-PO	30.2	60.4	50.1	**89.4**	58.3	**83.4**	17.1	28.0	20.1	32.0	18.1	53.3	34.0	**86.0**	40.2	**69.5**
+ LongFaith-PO	**60.5**	**66.4**	68.0	85.0	65.4	81.2	**38.1**	**30.5**	46.7	32.0	**50.2**	52.0	**73.7**	83.5	**55.6**	67.5

Table 2: Main experiment results on three multi-hop reasoning test sets and five long-context QA test sets from LongBench. The best results are in **bold** and second-best are underlined. **(S)** means single-doc QA task and **(M)** means multiple-docs QA task in LongBench. LongFaith-SFT and LongFaith-PO are synthesized by *GPT-4o mini* both in 2K size. To ensure fairness, we sample first 2K examples from baseline datasets.

4 Experiments

4.1 Implementation Details

Following previous studies, we leverage the training set of MuSiQue (Trivedi et al., 2022b), which is build on Wikipedia documents with supporting documents and correct answers. The officially retrieved 20 documents are provided and read only once in the input context in distractor setting. The statistics of training set is given in Tab. 6, covering *1K*, *2K*, *4K* and *8K*, where the balance of questions with different hops are considered. Following the pipeline we describe in Sec. 3, reasoning chains are samples to build LongFaith-SFT and LongFaith-PO. The statistics are presented in Tab. 1.

We conduct our experiments on a Linux server equipped with 4 A100-SXM4-40GB GPUs. For data synthesis of long-context reasoning instructions, we take *Llama-3.1-8B-Instruct* (Dubey et al., 2024), *Qwen2.5-7B-Instruct* (Yang et al., 2024), *Llama-3.1-70B-Instruct* and *GPT-4o mini* (Hurst et al., 2024) as generators and prompt LLMs to synthesize reasoning chains with vLLM (Kwon et al., 2023). We adopt the LoRA technique (Hu et al., 2021) for fine-tuning and ORPO technique (Hong et al., 2024) for preference optimization using the LLaMA-Factory framework (Zheng et al., 2024) to train *Llama-3.1-8B-Instruct*. Hyperparameters of post-training are given in App. F.

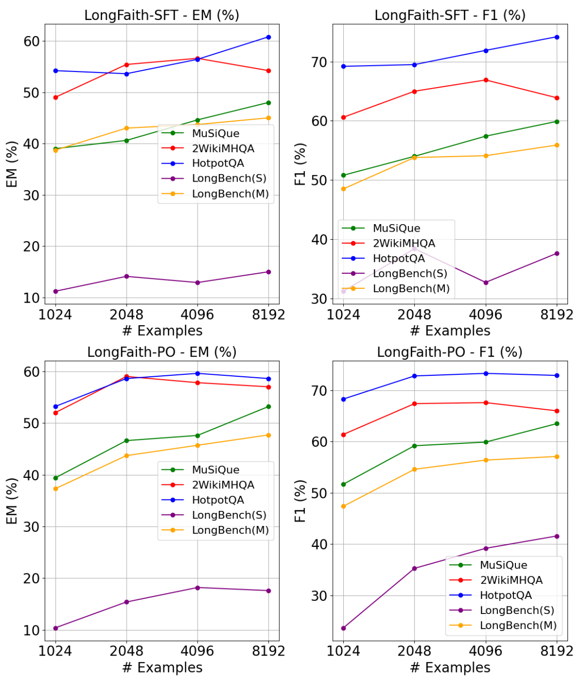

Figure 3: Performance of *Llama-3.1-8B-Instruct* trained on different size of instructions synthesized by *Qwen2.5-7B-Instruct* from *1K* to *8K*, evaluated by **EM** and **F1** metrics on multi-hop reasoning sets and LongBench.

4.2 Evaluation Setup

Following prior work (Li et al., 2024b), we utilize **three multi-hop reasoning datasets**, including MuSiQue (Trivedi et al., 2022b), 2WikiMultiHopQA (Ho et al., 2020), and HotpotQA (Yang et al., 2018), evaluating in distractor-setting, where the officially retrieved 10 or 20 documents are provided and read only once in the input context. We adopt the test sets sampled by (Trivedi et al., 2022a), with 500 examples in each set. Furthermore, in line with previous studies (Chen et al., 2024; Zhang et al., 2024c; Li et al., 2024a), we as-

Llama-3.1-8B-Instruct	MuSiQue				2WikiMultiHopQA			HotpotQA		
	Overall	2-Hop	3-Hop	4-Hop	Overall	2-Hop	4-Hop	Overall	Bridge	Comparison
Zero-Shot Prompting										
+ CoT	11.0	7.5	16.2	12.0	29.0	22.3	54.3	17.4	17.5	17.0
+ CoC	19.0	16.1	22.7	20.7	39.2	31.4	68.6	30.4	28.6	38.6
Superivised Fine-tuning										
+ LongFaith-SFT	40.6	44.1	37.7	35.9	55.4	51.1	71.4	53.6	57.0	37.5
w/o CoC	40.2	41.7	39.6	37.0	51.8	48.9	62.9	52.0	56.6	30.7
w/o GT	30.4	31.9	28.6	29.3	55.8	49.6	79.0	56.6	54.9	**64.8**
w/o Doc	20.0	23.6	18.2	13.0	55.8	47.1	**88.6**	47.4	45.9	54.5
Preference Optimization										
w/ GT-PO	44.0	45.7	42.9	41.3	56.0	50.4	77.1	54.4	58.3	36.4
w/ CoC-PO	43.6	44.5	44.8	39.1	53.2	48.6	70.5	56.2	59.2	42.0
w/ Doc-PO	41.4	42.5	40.3	40.2	56.0	52.7	68.6	56.4	59.5	42.0
+ LongFaith-PO	**46.6**	**47.2**	**48.1**	**42.4**	**59.0**	**55.9**	70.5	**58.6**	**59.7**	53.4

Table 3: Main experiment results on three long-context multi-hop reasoning datasets using the Exact-Match(**EM**) metric. The best results are in **bold**. The training set has 2K samples for both SFT and PO, synthesized by *Qwen2.5-7B-Instruct*. Detail statistics of synthetics datasets are presented in Tab. 1.

sess the performance on **LongBench** (Bai et al., 2023), which includes two test sets for single-doc QA including Qasper (Dasigi et al., 2021) and MultiFieldQA-EN (Bai et al., 2023), as well as three test sets for multi-docs QA tasks including HotpotQA, 2WikiMultiHopQA, and MuSiQue. Notably, although there is **an overlap in multi-hop reasoning tasks**, the LongBench version **further extends the lengths of document text**. To apply CoC prompting, single document is split into 20 even paragraphs with order. The statistics of datasets are listed in Tab. 7.

To ensure fairness, we use Substring Exact-Match (**SubEM**) (Yen et al., 2024; Li et al., 2024a) metric in main experiments, in case that models trained on baseline datasets are not good at instructions following to summarize the answer with "The answer is", and SubEM goes through the whole response to check whether the answer is in. Furthermore, following previous work (Choi et al., 2018; Zhang et al., 2024a; Li et al., 2024b), we use **EM** metric and **F1** scores for the trimmed part after "The answer is" for evaluation in main experiments and ablation studies. Comparing with LLM-as-a-Judge (Bai et al., 2024a; Chen et al., 2024; Zhang et al., 2024c) using strong API models like GPT-4o, the rule-based metrics cost much lower.

4.3 Baselines

We compete LongFaith-SFT and LongFaith-PO with datasets proposed in previous studies, including LongAlpaca (Chen et al., 2023b), LongAlign (Bai et al., 2024a), MuSiQue-Attribute (Li et al., 2024b), LongMIT (Chen et al., 2024), LongReward-SFT (Zhang et al., 2024c), SeaLong-SFT (Li et al., 2024a) for supervised fine-tuning, and LongReward-PO and SeaLong-PO for preference optimization. All of them aim at enhancing the performance of LLMs on long-context understanding, reasoning, and QA tasks. To ensure fairness, we keep the training setting consistent with App. F and truncate the size of training samples to a maximum of 2K.

4.4 Main Results

LongFaith Outperforms Previous Datasets. Following previous work and to ensure a fair comparison, we evaluate the performance of LongFaith on multi-hop reasoning test sets (Trivedi et al., 2022b; Ho et al., 2020; Yang et al., 2018) and LongBench (Bai et al., 2023), comparing it against baseline methods, including zero-shot prompting with Llama-3.1-8B-Instructand models trained on synthetic datasets from previous works. As shown in Tab. 2, LongFaith outperforms baseline datasets on most test sets. The performance of the model trained on LongFaith-PO surpasses that trained on LongFaith-SFT. This aligns with our expectations: compared to directly using positive samples for supervised fine-tuning, incorporating rejected samples to provide more fine-grained faithfulness preferences for optimization leads to better improvements in long-context reasoning and QA capabilities. We observe that some synthetic

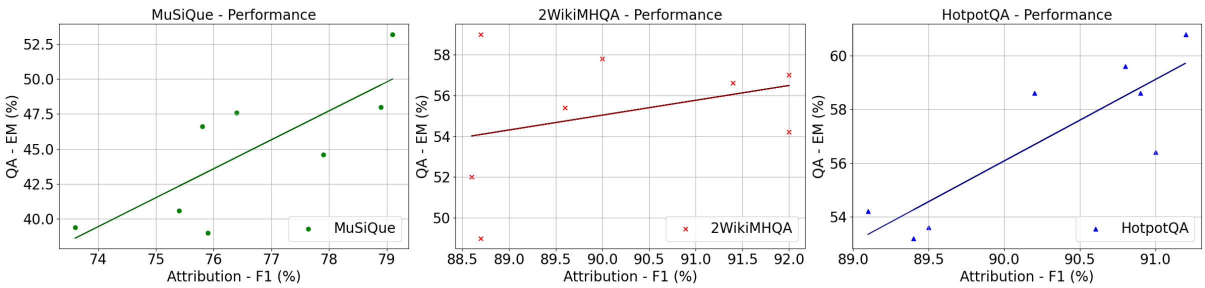

Figure 4: Scatter plot with a linear regression line fitting the relationship between **QA - EM** and **Attribution - F1** metrics on three long-context multi-hop reasoning test sets. A point refers to the performance of a model trained with a specific size between *1K* to *8K* by SFT or PO.

instruction sets degrade performance compared to native Llama-3.1-8B-Instruct. This proves that datasets with questionable faithfulness are even harmful to long-context reasoning ability of LLMs.

LONGFAITH Arrives at the Correct Answer without Redundant Exploration. We find that on 2WikiMultiHopQA, HotpotQA, and part of tasks in LongBench, SEALONG achieves a slight advantage in the SubEM metric against LONGFAITH, but fails in F1 scores. We investigate the length of response and present in Tab. 9. It turns out that the LLM trained on SEALONG conducts redundant exploration in response, producing more noisy content related to the answer, but actually arrives at a wrong answer, which means SubEM metric is easily to be "hacked". In contradiction, F1 scores requires to truncate the part after "The answer is", which demonstrates that a model trained on LONGFAITH datasets can arrive at the correct answer without redundant exploration and achieve a high score in a more strict metric. A case study is shown in Fig. 11 in Appendix.

Generalization. Based on statistics from Tab. 8, the main experiment demonstrates that LONGFAITH uses instructions with shorter context as input compared to baseline methods, reducing training costs while generalizing to LongBench tasks that require processing an average of 24K-70K tokens as input. This further highlights the generalization ability of our pipeline.

4.5 Analysis

Exploration on Different Perspective of Faithfulness. To validate the specific impact of different dimensions of faithfulness, we fine-tune models using negative samples as output and optimize using preference datasets that reject only a subset of negative samples. The statistics of the constructed

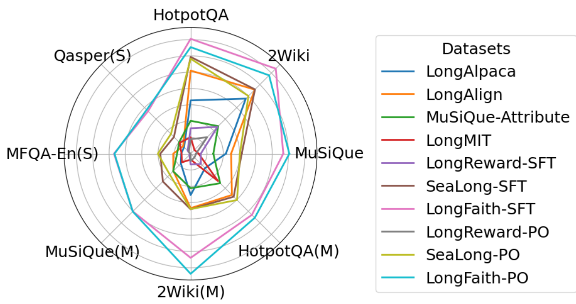

Figure 5: Visualization of F1 scores in Tab. 2.

datasets are shown in Tab. 1. Since each task in LongBench contains no more than 200 questions, performance evaluations can be prone to errors, so we chose to test on multi-hop reasoning datasets. Experimental results are shown in 3. The models trained with LONGFAITH-SFT and LONGFAITH-PO achieved high performance respectively in SFT and PO especially in F1 scores, as expected.

However, we note that in 4-Hop part of 2Wiki and comparison part of HotpotQA, LONGFAITH-SFT w/o CoC and w/o GT demonstrated better performance. Analysis reveals that for the question "Do both films, Cuban Colony and Prathyartha, have directors from the same country?", as the training set MuSiQue used specific entities as answers, the model responds "Both directors are from the same country, which is India. The answer is India.". Actually, the correct answer is "yes." Model trained on LONGFAITH w/o GT and LONGFAITH w/o Doc performed better with more exploration, but also lost overall performance due to hallucinations. Models trained on all PO datasets outperformed those trained using only positive samples for SFT, demonstrating the performance improvement brought by each fine-grained, credible preference. Finally, models trained on LONGFAITH-PO, which integrates three dimensions of faithfulness, achieved the best overall performance.

LLAMA-3.1-8B-INSTRUCT	MuSiQue		2WikiMHQA		HotpotQA		LongBench(S)		LongBench(M)	
+ LONGFAITH	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Superivised Fine-tuning										
w/ *Llama-3.1-8B-Instruct*	35.4	48.4	59.4	69.5	54.6	67.7	9.9	29.0	42.2	48.4
w/ *Qwen2.5-7B-Instruct*	40.6	54.0	55.4	65.0	53.6	69.5	14.1	38.4	43.0	53.8
w/ *Llama-3.1-70B-Instruct*	**44.8**	**58.1**	54.0	64.4	54.4	69.5	16.0	41.0	44.8	**56.7**
w/ *GPT-4o mini*	41.6	56.8	**64.6**	**73.8**	**55.4**	**70.5**	**16.9**	**42.0**	**47.2**	55.7
Preference Optimization										
w/ *Llama-3.1-8B-Instruct*	41.2	53.2	57.4	67.0	55.6	68.5	14.7	36.9	44.0	55.3
w/ *Qwen2.5-7B-Instruct*	46.6	59.2	59.0	67.4	**58.6**	**72.8**	15.4	35.3	44.8	55.6
w/ *Llama-3.1-70B-Instruct*	**50.4**	**63.2**	52.8	62.7	57.2	71.0	**16.4**	40.0	**48.3**	59.4
w/ *GPT-4o mini*	48.4	60.5	**59.8**	**68.0**	49.8	65.4	15.9	**42.4**	45.0	**59.8**

Table 4: Ablation study on various LLMs for synthesizing LONGFAITH-SFT and LONGFAITH-PO in the size of 2K. The base model for training and testing is *Llama-3.1-8B-Instruct*. **(S)** and **(M)** refer to Single-doc QA and Multi-docs QA in LongBench.

Scalability and Performance Gains. We explore the scaling-up potential of LONGFAITH on multi-hop reasoning test sets and LongBench. As presented in Tab. 7, we train *Llama-3.1-8B-Instruct* using LONGFAITH-SFT and LONGFAITH-PO synthesized by *Qwen2.5-7B-Instruct* across four dataset sizes, ranging from *1K* to *8K*. According to the performance trend in Fig. 3, LONGFAITH generally exhibits scaling-up potential, indicating that expanding the training dataset can further enhance performance. Moreover, LONGFAITH-PO, which incorporates fine-grained preference optimization, demonstrates a more stable upward trend compared to LONGFAITH-SFT, particularly in LongBench tasks. This result validates the robustness of the LONGFAITH pipeline.

Attribution-Based Reasoning Leads to Higher Performance. Utilizing CoC prompting for reasoning with attributions not only outperforms CoT in performance, as it presents in Tab. 3, but also provides greater interpretability and faithfulness as shown in Fig. 2. We use Attribution F1 as a metric to quantify the model's attribution capability using annotated supporting facts. Under CoC prompting, we analyze the references within reasoning chains, matching them against supporting facts like [1], [2], etc., and compute F1 scores based on recall and precision. We evaluate the attribution capability and overall performance of Llama-3.1-8B-Instructtrained on LONGFAITH-SFT and LONGFAITH-PO across four sizes and visualize the results in a scatter plot. The findings in 4 demonstrate a strong positive correlation between attribution capability and model performance, further validating the effectiveness of the LONGFAITH pipeline.

LLAMA-3.1-8B	LongBench(S)		LongBench(M)	
	EM	F1	EM	F1
w/ *10 docs*	11.2	31.7	39.7	52.2
w/ *20 docs*	15.4	35.3	44.8	55.6
w/ *30 docs*	16.0	36.0	46.1	56.7
w/ *40 docs*	16.9	37.8	47.2	58.1

Table 5: Effectiveness of LONGFAITH on LongBench as context length increases.

Impact of LLM Selection for Synthesis. We experimented with different LLMs for synthesis, including smaller open-source LLMs such as *Llama-3.1-8B-Instruct*, *Qwen2.5-7B-Instruct*, and larger open-source models like *Llama-3.1-70B-Instruct*, as well as a closed-source API model, *GPT-4o mini*, to synthesize LONGFAITH-SFT and LONGFAITH-PO for training Llama-3.1-8B-Instruct. The performance test results are presented in Tab.4. Using stronger closed-source API models to synthesize LONGFAITH-SFT led to a stronger performance boost, which aligns with intuition and previous work (Chen et al., 2024). However, an interesting finding is that the LONGFAITH-PO synthesized with different base LLMs did not show significant performance differences in preference optimization. Even smaller model like *Qwen2.5-7B-Instruct*, are able to synthesize high-quality reasoning chains, with performance on some datasets matching or even surpassing *GPT-4o mini*. This highlights the robustness of the LONGFAITH pipeline, which is capable of modeling fine-grained preferences to synthesize high-quality instructions.

Effectiveness of LONGFAITH as Context Length Increases. We further investigate the impact of increasing context length on the performance of models trained with the LongFaith pipeline. Specifically, we manipulate the number of documents included in the training set *MuSiQue* to simulate varying context lengths during training. By incrementally increasing the number of documents from 10 to 40, we assess how the model's reasoning ability scales when exposed to longer input contexts. The results, as reported in Tab. 5, show a consistent improvement in both EM and F1 scores across both LongBench(S) and LongBench(M) as the context length increases. This indicates that LongFaith effectively leverages additional contextual information, enhancing the model's comprehension and reasoning capabilities. These findings validate the scalability of the LongFaith framework in handling long-context scenarios, highlighting its potential for applications that require deep reasoning over extensive inputs.

5 Conclusion

This paper addresses the challenge of questionable faithfulness in data synthesis approaches for long-context LLMs. We propose LongFaith, a novel pipeline synthesizing faithful long-context reasoning datasets through ground truth integration and citation-based reasoning prompts. Experiments demonstrate its effectiveness, with ablation studies confirming the adaptability of the LongFaith-SFT and LongFaith-PO datasets across diverse long-context scenarios.

Limitations

While LongFaith demonstrates significant improvements in long-context reasoning tasks, its scalability and generalization to other LLMs remain an open question. Our experiments focused on a single model, and thus, the performance of LongFaith on other general-purpose LLMs still needs further validation. Additionally, while the synthesized instruction sets with lengths of approximately 10,000 tokens successfully generalized to long-context reasoning tasks, future work will explore the extension of LongFaith to generate instructions with even longer contexts and evaluate the impact on model performance. Finally, LongFaith currently concentrates on reasoning tasks, and we plan to explore its generalization to other tasks such as summarization, dialogue generation, and others, to assess its broader applicability.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. Make your llm fully utilize the context. *arXiv preprint arXiv:2404.16811*.

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. *arXiv preprint arXiv:2401.18058*.

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, et al. 2024b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

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

Moshe Berchansky, Daniel Fleischer, Moshe Wasserblat, and Peter Izsak. 2024. Cotar: Chain-of-thought attribution reasoning with multi-level granularity. *arXiv preprint arXiv:2404.10513*.

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

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023b. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

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

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

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. *arXiv preprint arXiv:2105.03011*.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*.

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. *arXiv preprint arXiv:2402.13753*.

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*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *International Conference on Machine Learning*.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. *arXiv preprint arXiv:2404.03381*.

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. *arXiv preprint arXiv:2402.10171*.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2024. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2024. 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.

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*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

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, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, et al. 2024. Learning fine-grained grounded citations for attributed large language models. *arXiv preprint arXiv:2408.04568*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Xuhui Jiang, Shengjie Ma, Chengjin Xu, Cehao Yang, Liyu Zhang, and Jian Guo. 2025. Synthesize-on-graph: Knowledgeable synthetic data generation for continue pre-training of large language models. *arXiv preprint arXiv:2505.00979*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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*.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. *arXiv preprint arXiv:2311.03731*.

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*.

Yanyang Li, Shuo Liang, Michael R Lyu, and Liwei Wang. 2024b. Making long-context language models better multi-hop reasoners. *arXiv preprint arXiv:2408.03246*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. 2024. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. *arXiv preprint arXiv:2407.10805*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv: 1707.06347*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph. *arXiv preprint arXiv:2307.07697*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022a. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

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.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*.

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

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. 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*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Liu Yong, and Shen Huang. 2024a. End-to-end beam retrieval for multi-hop question answering. 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 1718–1731.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, et al. 2024b. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv preprint arXiv:2409.02897*.

Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024c. Longreward: Improving long-context large language models with ai feedback. *arXiv preprint arXiv:2410.21252*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. 2024d. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

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. *arXiv preprint arXiv:2309.10400*.

A An Example of Synthesized LongFaith-SFT and LongFaith-PO Datasets

Construction of LongFaith-SFT and LongFaith-PO using Chosen Sample and Rejected Samples

LongFaith -SFT

Instruction

You are provided with documents, a complex logical reasoning question, and the correct answer.
You must refer to the documents to perform step-by-step logical reasoning and reach the correct answer.
Each reasoning step must be on a separate line, ending with a newline character.
Cite the document properly during reasoning, e.g., `[1]`, `[2]`, etc.
End your reasoning with `The answer is` followed by the correct answer.

DOCUMENTS:
[1] Muammar Gaddafi: From childhood, Gaddafi was aware of the involvement of European colonialists in Libya……
[2] Muammar Gaddafi: On 16 April 1973, Gaddafi proclaimed the start of a "Popular Revolution" in a Zuwarah speech……
……
[9] Bogotá: Bogotá (/ ˈboʊɡətɑː /, / ˌbɒɡəˈtɑː /, / ˌboʊ - /; Spanish pronunciation: (boɣoˈta) (listen))……
[10] Alanje: Alanje is a corregimiento in Alanje District, Chiriquí Province, Panama. It is the seat of Alanje District……
……
[20] All Saints Church, Lockerbie: All Saints Church is in Ashgrove Terrace, Lockerbie, Dumfries and Galloway, Scotland……

QUESTION:
When was the country containing Alanje colonized by the country housing the All Saints Church in the town known for the bombing by Gaddafi's Libya?

Chosen Output (Sample: Highest Faithfulness)

Step 1: Identify the country where Alanje is located.
From Document [10], Alanje is in the Alanje District, Chiriquí Province, Panama. Thus, Alanje is located in Panama.

Step 2: Identify the country where All Saints Church is located.
……
The answer is 1698.

LongFaith -PO

Rejected Output (Sample: Misinformation due to Lack of Verification)

The All Saints Church is located in Lockerbie, Scotland, which is part of the United Kingdom.

The UK was involved in colonizing various territories, including Panama, where Alanje is located. [10]
……

The answer is 1903.

Rejected Output (Sample: Reasoning without Attribution)

Step 1: Identify the country containing Alanje.
From the document on Alanje, it is stated that Alanje is located in Panama.

Step 2: Identify the country housing the All Saints Church in Lockerbie.
……
The answer is 1698.

Rejected Output (Sample: Potential Knowledge Conflicts)

Alanje is a location, so we need to determine which country contains Alanje. Alanje is located in Panama.
……
However, Panama was not colonized by the United Kingdom; Panama was colonized by Spain.
So, we need to determine the connection between the United Kingdom and Spain regarding the colonization of Panama.
……
The answer is 1698.

Figure 6: An Example of synthesized LongFaith-SFT and LongFaith-PO datasets.

B Prompt Templates

We present the prompt templates that are used to synthesize the datasets. The core prompt template that generates long-context reasoning chains guided by ground truth using chain-of-citation is shown in Fig. 7. The samples are used in LONGFAITH-SFT dataset and are chosen as positive in LONGFAITH-PO dataset, since they are of the highest faithfulness. The other three prompt templates synthesize rejected samples for LONGFAITH-PO dataset, corresponding to (1) Misinformation due to lack of verification (Fig. 8), (2) Reasoning without attribution (Fig. 9), and (3) Potential knowledge conflicts (Fig. 10).

Synthesis of Long-Context Reasoning Chains Guided by Ground Truth Using Chain-of-Citation

Prompt for Synthesis

You are provided with documents, a complex logical reasoning question, and the correct answer.
You must refer to the documents to perform step-by-step logical reasoning and reach the correct answer.
Each reasoning step must be on a separate line, ending with a newline character.
Cite the document properly during reasoning, e.g., `[1]`, `[2]`, etc.
End your reasoning with `The answer is` followed by the correct answer.

DOCUMENTS:
[10] Alanje: Alanje is a corregimiento in Alanje District, Chiriquí Province, Panama……..
[14] British Empire: In 1695, the Scottish Parliament granted a charter to the Company of Scotland……
[16] Muammar Gaddafi: In 1977, Gaddafi dissolved the Republic and created a new socialist state, ……
[20] All Saints Church, Lockerbie: All Saints Church is in Ashgrove Terrace, Lockerbie, ……

QUESTION:
When was the country containing Alanje colonized by the country housing the All Saints Church in the town known for the bombing by Gaddafi's Libya?

ANSWER:
1698

Synthesized Reasoning Chains

Step 1: Identify the country where Alanje is located.
From Document [10], Alanje is in the Alanje District, Chiriquí Province, Panama. Thus, Alanje is located in Panama.

Step 2: Identify the country where All Saints Church is located.
From Document [20], All Saints Church is in Lockerbie, Dumfries and Galloway, Scotland. Thus, All Saints Church is in Scotland.

Step 3: Determine the historical event that connects Panama and Scotland.
From Document [14], in 1698, the Company of Scotland established a settlement on the isthmus of Panama. This settlement was part of the Darien scheme, a Scottish colonial venture.

Step 4: Confirm the timeline of the colonization.
From Document [14], the Scottish settlement on the isthmus of Panama occurred in 1698.

Step 5: Verify the connection to Gaddafi's Libya and Lockerbie.
From Document [16], Gaddafi's Libya was responsible for the Lockerbie bombing. This connects the country housing All Saints Church (Scotland) to the logical context of the question.

Step 6: Synthesize the information.
The country containing Alanje (Panama) was colonized by the country housing All Saints Church (Scotland) in 1698.

The answer is 1698.

Chosen as Positive Sample: Highest Faithfulness

Figure 7: An example of synthesized chosen reasoning chain. The current reasoning chain must arrive at a correct answer, and reasoning with proper citation proposes more faithfulness and interpretability. Therefore, LONGFAITH will choose it in supervised fine-tuning and preference optimization as positive sample.

Synthesis of Long-Context Reasoning Chains Using Chain-of-Citation

Prompt for Synthesis

You are provided with documents, a complex logical reasoning question, and the correct answer.
You must refer to the documents to perform step-by-step logical reasoning and reach the correct answer.
Each reasoning step must be on a separate line, ending with a newline character.
Cite the document properly during reasoning, e.g., `[1]`, `[2]`, etc.
End your reasoning with `The answer is` followed by the correct answer.

DOCUMENTS:
[1] Muammar Gaddafi: From childhood, Gaddafi was aware of the involvement of European colonialists in Libya……
[2] Muammar Gaddafi: On 16 April 1973, Gaddafi proclaimed the start of a "Popular Revolution" in a Zuwarah speech……
……
[9] Bogotá: Bogotá (/ ˈboʊɡətɑː /, / ˌbɒɡəˈtɑː /, / ˌboʊ - /; Spanish pronunciation: (boɣoˈta) (listen))……
[10] Alanje: Alanje is a corregimiento in Alanje District, Chiriquí Province, Panama. It is the seat of Alanje District……
……
[20] All Saints Church, Lockerbie: All Saints Church is in Ashgrove Terrace, Lockerbie, Dumfries and Galloway, Scotland……

QUESTION:
When was the country containing Alanje colonized by the country housing the All Saints Church in the town known for the bombing by Gaddafi's Libya?

Synthesized Reasoning Chains

The All Saints Church is located in Lockerbie, Scotland, which is part of the United Kingdom.

The UK was involved in colonizing various territories, including Panama, where Alanje is located. [10]

The British Empire's colonization of Panama through the Darien scheme began in 1698, although this attempt was unsuccessful. [14]

The formal British colonization of Panama occurred later, after the separation of Panama from Colombia in 1903, when the UK recognized Panama's independence.

The answer is 1903.

Chosen as Rejected Sample: Misinformation due to Lack of Verification

Figure 8: An example of synthesized rejected reasoning chain. Misinformation due to lack of verification will cause more hallucination if we use current reasoning chain to fine-tune a LLM. Therefore, LONGFAITH will reject it in preference optimization.

Synthesis of Long-Context Reasoning Chains Guided by Ground Truth Using Chain-of-Thought

Prompt for Synthesis

You are provided with documents, a complex logical reasoning question, and the correct answer.
You must refer to the documents to perform step-by-step logical reasoning and reach the correct answer.
Each reasoning step must be on a separate line, ending with a newline character.
End your reasoning with `The answer is` followed by the correct answer.

DOCUMENTS:
Alanje: Alanje is a corregimiento in Alanje District, Chiriquí Province, Panama……..
British Empire: In 1695, the Scottish Parliament granted a charter to the Company of Scotland……
Muammar Gaddafi: In 1977, Gaddafi dissolved the Republic and created a new socialist state, ……
All Saints Church, Lockerbie: All Saints Church is in Ashgrove Terrace, Lockerbie, ……

QUESTION:
When was the country containing Alanje colonized by the country housing the All Saints Church in the town known for the bombing by Gaddafi's Libya?

ANSWER:
1698

Synthesized Reasoning Chains

Step 1: Identify the country containing Alanje.
From the document on Alanje, it is stated that Alanje is located in Panama.

Step 2: Identify the country housing the All Saints Church in Lockerbie.
The document on All Saints Church specifies that it is in Lockerbie, Scotland.

Step 3: Determine the relationship between Panama and Scotland.
The document on the British Empire mentions that in 1698, the Scottish Parliament established a settlement on the isthmus of Panama.

Step 4: Verify if the settlement was during a colonization attempt.
The British Empire document states that the Scottish settlement on the isthmus of Panama was part of the Darien scheme, an effort by Scotland to establish its own overseas empire, which qualifies as a colonization attempt.

Step 5: Confirm the year of colonization.
The British Empire document explicitly states that the settlement in Panama by Scotland occurred in 1698.

The answer is 1698.

Chosen as Rejected Sample: Reasoning without Attribution

Figure 9: An example of synthesized rejected reasoning chain. As it mentioned in previous work, lack of attribution will lead to much more interpretability and faithfulness, and response with citation is encouraged. Therefore, LONGFAITH will reject it in preference optimization.

Synthesis of Long-Context Reasoning Chains Guided by Ground Truth Without Documents Using CoT

Prompt for Synthesis

You are provided with a complex logical reasoning question and the correct answer.
You must perform step-by-step logical reasoning and reach the correct answer.
Each reasoning step must be on a separate line, ending with a newline character.
End your reasoning with `The answer is` followed by the correct answer.

QUESTION:
When was the country containing Alanje colonized by the country housing the All Saints Church in the town known for the bombing by Gaddafi's Libya?

ANSWER:
1698

Synthesized Reasoning Chains

Let's break this down step by step:

Alanje is a location, so we need to determine which country contains Alanje. Alanje is located in Panama.

Next, we need to identify the country housing the All Saints Church in the town known for the bombing by Gaddafi's Libya.

The bombing by Gaddafi's Libya refers to an event in 1986 in Loughborough, England.

Loughborough, England, is in the United Kingdom, and it indeed has an All Saints Church.

Now we know we are looking for when Panama (containing Alanje) was colonized by the United Kingdom (housing the All Saints Church in Loughborough).

However, Panama was not colonized by the United Kingdom; Panama was colonized by Spain.

So, we need to determine the connection between the United Kingdom and Spain regarding the colonization of Panama.

The United Kingdom is not directly related to the colonization of Panama, so the question appears to involve a broader historical context.

Spain colonized Panama in 1501, but colonization efforts and territorial control fluctuated.

The specific colonization of Alanje, a town in Panama, is tied to the Spanish colonization of the region.

The date 1698 aligns with significant historical colonization activities in the Americas.

Thus, the colonization being referred to in the question likely took place in 1698.

The answer is 1698.

Chosen as Rejected Sample: Potential Knowledge Conflicts

Figure 10: An example of synthesized rejected reasoning chain. Information related to "Panama was colonized by Spain" is not mentioned in documents provided, which will cause knowledge conflicts if using current reasoning chain for fine-tuning. Therefore, LONGFAITH will reject it in preference optimization.

C Case Study of Metric "Hack" on SubEM comparing previous study with LongFaith

Who was the first elected president of the country where, along with the nation where Ágnes Konkoly is from, many expelled French Jews relocated to?

[1] Middle Ages: Jewish communities were expelled from England in 1290 and from France in 1306. Although some……

[2] Konkoly: Ágnes Konkoly (born 23 July 1987) is a Hungarian model, wedding planner and beauty pageant……

Correct Answer: **Gabriel Narutowicz**

SeaLong

To find the answer, we need to identify the country where Ágnes Konkoly is from and the country where many expelled French Jews relocated to.

Step 1: Identify the country where Ágnes Konkoly is from.
Ágnes Konkoly is a Hungarian model, wedding planner, and beauty pageant titleholder, so she is from Hungary.

Step 2: Identify the country where many expelled French Jews relocated to.
From the given documents, we can see that many expelled French Jews relocated to Poland and Hungary.

Step 3: Find the first elected president of the country where Ágnes Konkoly is from (Hungary).
The first president of Hungary was Lajos Kossuth, but he was not elected president. The first freely elected president of Hungary was János Áder.

However, the problem asks for the first elected president of the country where Ágnes Konkoly is from, which is Hungary. The first freely elected president of Hungary was János Áder, but the first president of the country was Lajos Kossuth, who was not elected.

Step 4: Determine the first elected president of the country where many expelled French Jews relocated to (Poland).
The first president of Poland was **Gabriel Narutowicz**, who was elected in 1922.

However, the problem asks for the first elected president of the country where Ágnes Konkoly is from, which is Hungary. We already found that the first freely elected president of Hungary was János Áder.

The answer is **János Áder**.

LongFaith

To determine the first elected president of the country where many expelled French Jews relocated to, we need to analyze the information provided in the documents.

First, we identify the country where Ágnes Konkoly is from. Ágnes Konkoly is a Hungarian model, and Hungary is her country of origin [18].

Next, we look for information regarding the relocation of French Jews. Document [19] states that after France expelled its original Jewish population in the Middle Ages, many of these Jews settled in areas that are now part of Germany, specifically in Alsace, which was formerly part of Germany.

Now, we need to find the first elected president of Poland, as Poland is the country where many of these French Jews relocated. According to document [9], the first president of Poland was Gabriel Narutowicz, who was sworn in on 11 December 1922.

Thus, the first elected president of Poland, where many expelled French Jews relocated, is indeed Gabriel Narutowicz.

The answer is **Gabriel Narutowicz**.

Figure 11: A case study that SubEM metric is "hacked" by previous study, which conduct more exploration with redundancy in response. LongFaith can arrive at the final correct answer with shorter response.

D Post-Training

In this section, we present two post-training algorithms—Supervised Fine-Tuning (SFT) and Preference Optimization (PO)—to better leverage synthetic data for efficiently enhancing model performance. Specifically, the model performs supervised fine-tuning on high-quality faithful outputs or is trained through reinforcement learning using synthetic preference pairs.

Supervised Fine-tuning on Faithful Outputs We minimize the negative log-likelihood of the output as follows:

$$\begin{aligned}\mathcal{L}_{\text{SFT}} &= -\frac{1}{|y|}\log \pi_\theta(y \mid x) \\ &= -\frac{1}{|y|}\sum_{i=1}^{|y|}\log \pi_\theta(y_i \mid x, y_{<i})\end{aligned} \tag{7}$$

where y denotes the high-quality faithful outputs, which are synthesized in Section 3.

Reinforcement Learning from Synthetic Preference Additionally, we can leverage synthetic preference pairs for reinforcement learning (RL) to fine-tune the model toward generating faithful outputs while reducing the likelihood of low-scoring outputs. Standard RL algorithms for optimizing LLMs include Proximal Policy Optimization (PPO) (Schulman et al., 2017), RLOO (Ahmadian et al., 2024). However, these methods incur high computational costs. Recent approaches such as Direct Preference Optimization (DPO) (Rafailov et al., 2024), Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024), and Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024) have been proposed to mitigate both computational and data requirements. In this work, we adopt the ORPO algorithm, which achieves an optimal balance between computational efficiency and model performance.

ORPO introduces an odds ratio loss $\mathcal{L}_{OR}$ that minimizes the negative log odds ratio between preferred ("win" y_w) and dispreferred ("lose" y_l) outputs:

$$\mathcal{L}_{\text{OR}} = -\log\sigma\left(\log\frac{\text{odds}_\theta(y_w|x)}{\text{odds}_\theta(y_l|x)}\right) \tag{8}$$

where σ denotes the sigmoid function and $\text{odds}_\theta(y|x) = \frac{\pi_\theta(y|x)}{1-\pi_\theta(y|x)}$ measures how much more likely y is to be generated. The final objective of ORPO is to combine the SFT loss and the OR loss while controlling their relative importance through a hyperparameter β:

$$\mathcal{L}_{\text{ORPO}} = \mathcal{L}_{\text{SFT}} + \beta \cdot \mathcal{L}_{\text{OR}} \tag{9}$$

In this paper, the chosen output y_w is synthesized by LongFaith through comprehensive consideration of Supporting Docs, Chain-of-Citation (CoC), and Ground Truth (GT), and is consequently assigned a high score. Conversely, the rejected output y_l refers to synthesized outputs that lack at least one of these three critical elements (Supporting Docs, CoC, or GT), which are deemed low-scoring due to insufficient design considerations.

E Statistics of Main Experiments

MuSiQue	#2-Hop	#3-Hop	#4-Hop
1K	0	512	512
2K	512	512	1024
4K	1024	2048	1024
8K	3072	4096	1024

Table 6: Statistics of train set for synthesis in different size sampled from **MuSiQue** (Trivedi et al., 2022b).

Datasets	#Count	Avg. L.	Max L.
Multi-Hop Reasoning			
MuSiQue	500		
2-Hop	254	10843.3	17560
3-Hop	154	11456.5	19225
4-Hop	92	11224.3	16756
2WikiMultiHopQA	500		
2-Hop	395	4449.5	10631
4-Hop	105	4041.4	9365
HotpotQA	500		
Bridge	412	6301.0	15702
Comparison	88	5777.6	11939
LongBench			
Qasper (S)	200	24262.3	101636
MultiFieldQA-En (S)	150	29583.7	64751
MuSiQue (M)	200	69876.8	82338
2WikiMHQA (M)	200	30076.5	72971
HotpotQA (M)	200	57041.4	81815

Table 7: Statistics of test sets including three long-context multi-hop reasoning datasets sampled by (Trivedi et al., 2022a) and five long-context QA datasets from LongBench (Bai et al., 2023). **Avg. L.** and **Max L.** refer to the average length and max length of input prompts for test samples. **(S)** and **(M)** refer to Single-doc QA and Multi-doc QA in LongBench.

Datasets	Instruction	Output(Chosen)	Rejected
LongAlpaca	52043.2	620.7	0
LongAlign	36307.2	1412.6	0
MuSiQue-Attribute	11395.0	343.7	0
LongMIT	280808.9	825.2	0
LongReward	72892.2	913.4	960.6
SEALONG	82248.6	1156.5	1139.1
LONGFAITH	11542.1	1029.6	896.7

Table 8: Average text length of baseline datasets and LONGFAITH in main experiments in Tab. 2. All of them has *2K* examples.

Datasets	MuSiQue	2WikiMultiHopQA	HotpotQA	Qasper	MultiFieldQA	Avg.L
LongAlpaca	365.62	372.97	319.60	657.34	511.25	445.36
LongAlign	493.56	349.65	371.77	651.76	623.15	497.98
MuSiQue-Attribute	99.61	168.74	164.24	317.75	252.95	200.66
LongMIT	138.03	159.16	116.40	194.41	196.43	160.89
LongReward-SFT	285.20	241.47	178.26	750.83	537.95	398.74
SeaLong-SFT	1091.54	776.01	926.29	1035.77	822.82	930.49
LongFaith-SFT	820.04	619.18	771.68	1056.55	941.13	841.72
LongReward-PO	219.90	253.41	179.44	616.11	460.40	345.85
SeaLong-PO	961.51	740.14	891.75	946.68	826.77	873.37
LongFaith-PO	831.17	669.76	786.77	1034.71	917.11	847.90

Table 9: Average length of model output in test sets trained on different synthesized instruction.

F Hyperparameters

Hyperparameters	Value
# GPUs used	4
Learning rate	5e-5
Per-device batch size	1
Gradient accumulation steps	8
LoRA rank	32
LoRA alpha	64
LoRA dropout	0.1
ORPO beta	0.1
Warm-up ratio	0.1
Epochs	1
Precision	bfloat16
Optimizer	AdamW

Table 10: Hyperparameter settings of fine-tuning and preference optimization.