Making Long-Context Language Models Better Multi-Hop Reasoners

Yanyang Li[1], **Shuo Liang**[1,2], **Michael R. Lyu**[1], **Liwei Wang**[1*]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]Shanghai AI Laboratory
{yyli21,sliang23,lyu,lwwang}@cse.cuhk.edu.hk

Abstract

Recent advancements in long-context modeling have enhanced language models (LMs) for complex tasks across multiple NLP applications. Despite this progress, we find that these models struggle with multi-hop reasoning and exhibit decreased performance in the presence of noisy contexts. In this paper, we introduce *Reasoning with Attributions*, a novel approach that prompts LMs to supply attributions for each assertion during their reasoning. We validate our approach through experiments on three multi-hop datasets, employing both proprietary and open-source models, and demonstrate its efficacy and resilience. Furthermore, we explore methods to augment reasoning capabilities via fine-tuning and offer an attribution-annotated dataset and a specialized training strategy. Our fine-tuned model achieves competitive performance on multi-hop reasoning benchmarks, closely paralleling proprietary LMs such as ChatGPT and Claude-instant[1].

1 Introduction

The field of long-context modeling has garnered significant attention due to its importance in applications that demand extensive comprehension and generation capabilities (Lewis et al., 2020; Liu et al., 2023b). Techniques for long-context modeling (Chen et al., 2023a; Peng et al., 2023; Chen et al., 2023b) have been proposed with encouraging results on established benchmarks (An et al., 2023; Bai et al., 2023).

Nevertheless, we have identified a gap in the performance of these models when it comes to multi-hop reasoning tasks, where a model must navigate and synthesize information from disparate sources to answer complex questions. Evidence from key benchmarks such as LongBench (Bai et al., 2023), as well as our experimental results in Section 4.3, indicate that these long-context LMs underperform compared to leading multi-hop reasoning systems (Zhang et al., 2023). The reasons for this shortfall in multi-hop reasoning effectiveness are not yet fully understood.

We contend that the limitations in multi-hop reasoning observed in long-context LMs stem from two main issues: The inability to discern pertinent information within noisy contexts (Liu et al., 2023a) and the struggle to incorporate knowledge within the context effectively, particularly for smaller-scale models (Zheng et al., 2023a). To address these challenges, we introduce *Reasoning with Attributions*, a methodology that compels LMs to substantiate their reasoning by linking assertions to relevant context segments, such as citations (Gao et al., 2023) or direct quotations (Menick et al., 2022). This approach not only guides LMs to perform targeted information retrieval to identify the position of relevant contexts, thereby reducing noise, but also ensures their responses are well-grounded in the source material. Our preliminary and comprehensive experimental findings, detailed in Sections 2.1 and 4.3, confirm the efficacy and resilience of this method across various multi-hop reasoning benchmarks.

Despite these advancements, smaller long-context LMs exhibit continued difficulties in reasoning. We explore the potential for these models to improve through learning to reason and attribute simultaneously. Utilizing ChatGPT (Brown et al., 2020) to annotate the multi-hop reasoning dataset MuSiQue (Trivedi et al., 2022), we create a specialized dataset *MuSiQue-Attribute* for fine-tuning models in this dual capacity. We propose a potent learning strategy that leverages multi-task learning and data augmentation to fully exploit these annotations. Our experiments with five long-context LMs across three multi-hop reasoning datasets and two general instruction-following datasets reveal

*Corresponding author.
[1]The dataset, model, and code are publicly available at https://github.com/LaVi-Lab/LongContextReasoner.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2462–2475

that our fine-tuned Vicuna-7B model (Zheng et al., 2023b) surpasses similar-scale baselines by a substantial margin, i.e., more than 20 points on average, and even outperforms ChatGPT and Claude-instant on MuSiQue, albeit with a slight trade-off in other capabilities. This study illuminates a promising avenue to enhance the multi-hop reasoning capabilities of long-context LMs through a focus on attributions.

Our contributions can be summarized as follows:

- We introduce *Reasoning with Attributions*, an innovative reasoning paradigm that enhances both the performance and robustness of long-context LMs in multi-hop reasoning tasks.
- We provide *MuSiQue-Attribute*, an attribution-annotated multi-hop reasoning dataset to support further research in this domain.
- We develop a bespoke learning strategy that incorporates novel auxiliary tasks for multi-task learning and employs tailored data augmentation techniques.
- Through rigorous testing on three multi-hop reasoning datasets and two general instruction-following benchmarks with both proprietary and open-source models, we demonstrate that our fine-tuned Vicuna-7B model achieves comparable multi-hop reasoning performance to ChatGPT with minimal impact on other capabilities. Additionally, our model shows resilience against varying degrees of contextual noise, underscoring the effectiveness of our methods.

2 Reasoning with Attributions

2.1 Pilot Study

The challenge of large language models becoming mired in irrelevant contexts, known as the "Lost in the Middle" phenomenon, has been documented across various NLP tasks, such as multi-document QA (Liu et al., 2023a) and mathematical reasoning (Shi et al., 2023). This issue is also apparent in multi-hop reasoning, which we illustrate later in Figure 2. Prior research has noted this problem but has not decoded the underlying mechanisms. For example, while Liu et al. (2023a) found that introducing the query before the context can aid in better information retrieval from the context, they did not achieve an improvement in QA performance using this query-aware approach. As we suggest in Section 1, the reasons might extend beyond mere retrieval challenges to include complications in effectively applying the retrieved knowledge.

Instruction: Write an accurate and concise answer for ...   <Retrieve for the question>   Document [1](Title: David Myles (musician)): ...   Document [2](Title: Jamal Plays Jamal): ...   Document [3](Title: Top and Bottom Brass): ...   (Other retrieved documents are omitted.)   **Question**: What is the genre of the record label of the band that performed on the Crush Tour?   **Answer**:
CoT: The Crush Tour is performed by the band Bon Jovi. The record label of Bon Jovi is Island Records. The genre of Island Records is jazz. The answer is: jazz ✔
CoC: The Crush Tour is performed by the band Bon Jovi [8]. The record label of Bon Jovi is Island Records [17]. The genre of Island Records is jazz [19]. The answer is: jazz ✔
CoQ: The Crush Tour is performed by the band Bon Jovi ("The Crush Tour is a third concert" [8]). The record label of Bon Jovi is Island Records ("Bounce is the eighth studio album by American" [17]). The genre of Island Records is jazz ("The Antidote is the debut album by English jazz" [19]). The answer is: jazz ✔

Table 1: An example of CoT and two of our reasoning with attribution methods: CoC and CoQ. We highlight the differences between these methods, e.g., answers are marked in green, citations are marked in orange and quotes are marked in blue.

To tackle the issues outlined earlier, we introduce *Reasoning with Attributions*, a strategy that mandates language models to link the claims made during reasoning to specific sections of the provided context. This implicit requirement effectively decomposes a complex multi-hop question into two more manageable tasks: Pinpointing pertinent information within the context and constructing well-founded claims based on that information.

We adapt the concept of Chain-of-Thought (CoT) (Wei et al., 2022) reasoning to create two distinct variants aligned with our attribution-based approach: **Chain-of-Citation** (CoC) and **Chain-of-Quote** (CoQ). In CoC, models are prompted to reference citations corresponding to each step of the reasoning chain. CoQ goes further by requiring models to include direct quotations from the cited material for each reasoning step. An illustrative example highlighting the nuances between these methods is provided in Table 1.

The results of our preliminary study (Please refer to Section 4 for the setup), detailed in Ta-

Model	MuSiQue		2Wiki		HotpotQA	
	EM	F1	EM	F1	EM	F1
ChatGPT (`gpt-3.5-turbo-1106`)						
+ AO	15.8	26.9	46.2	57.2	51.0	65.4
+ CoT	36.2	50.1	55.2	70.1	56.8	71.2
+ CoC	**37.0**	51.0	**55.4**	**71.1**	**58.6**	**73.4**
+ CoQ	36.4	**51.3**	54.0	68.7	55.4	70.2
Claude-instant (`claude-instant-1.2`)						
+ AO	26.2	39.4	47.0	57.5	**54.4**	**68.4**
+ CoT	26.0	37.9	40.8	52.3	20.2	26.3
+ CoC	**32.2**	**46.2**	**53.4**	**67.0**	54.2	68.3
+ CoQ	30.2	45.9	49.8	62.1	50.8	65.0

Table 2: Exact-Match (EM) and F1 scores of ChatGPT and Claude-instant with 5-shot prompting on multi-hop reasoning datasets, e.g., MuSiQue, 2WikiMultiHopQA (2Wiki for short) and HotpotQA. The best results are in **bold**. AO means models predict answers only.

Error Type	Portion
Incorrect Answer	58.44%
Non-Existent Attributions	12.56%
Incorrect Citations	9.80%
Repeated Citations	6.35%
Extreme Quotes	10.55%

Table 3: Incidence rates of different error types.

Entry	Value
#Max Words per Sample	3385
#Mean Words per Sample	1809.10
#Averaged Words per CoT Step	11.64
#Averaged Words per Quote	16.60
#Totoal Samples	1358
2-Hop Samples [%]	82.18%
3-Hop Samples [%]	14.06%
4-Hop Samples [%]	3.76%

Table 4: Statistics of MuSiQue-Attribute.

ble 2, compare the efficacy of CoT, CoC, and CoQ when applied to two proprietary long-context LMs: ChatGPT (Brown et al., 2020) and Claude-instant (Bai et al., 2022). Without further notice, ChatGPT always refers to `gpt-3.5-turbo-1106` and `claude-instant-1.2` for Claude-instant in this work. The findings suggest that both CoC and CoQ generally yield improvements over CoT, indicating that attribution-based reasoning enhances the precision and coherence of the models' reasoning processes. CoQ appears to slightly underperform CoC, likely due to the increased complexity of producing exact quotations.

It is noteworthy that even in instances where CoT reduces the Answer Only (AO) performance, CoC is able to not only mitigate this decline but also surpass the AO baseline. This demonstrates the potential of CoC as a robust reasoning method. The success of our approach with various open-sourced models is further elaborated upon in Section 4.3. Based on these insights, we adopt CoC as our primary reasoning format in subsequent sections.

2.2 Dataset Curation

Our analysis, evidenced by the data in Tables 2 and 5, confirms that while reasoning with attributions holds promise, smaller open-source long-context language models significantly underperform compared to their proprietary counterparts in multi-hop reasoning tasks. To address this, we investigate whether training these models to perform attributions can boost their reasoning capabilities.

A hurdle in this process is the lack of attribution annotations within existing multi-hop reasoning benchmarks. To bridge this gap, we have generated new annotations by prompting ChatGPT with 5-shot CoQ. This has been done to create CoT with attributions for 5,000 instances randomly selected from the answerable training set of the MuSiQue dataset (Trivedi et al., 2022). Although CoC generally outperforms CoQ, we chose CoQ for annotation because it provides more detailed information. This richness is beneficial not only for evaluating the quality of the annotations but also proves advantageous for the fine-tuning processes discussed in Section 3.

After generating the annotations, we implemented a filtering process to exclude annotations with any of the following errors (Please refer to Appendix E for implementation details):

- **Incorrect Answer**: The model's predicted answer does not align with the reference answer, which typically indicates an erroneous CoT.
- **Non-Existent Attributions**: Fabricated citations or quotes that do not correspond to the actual context are indicative of model hallucination.
- **Incorrect Citations**: Citations do not match the manually identified supporting facts, suggesting flawed attributions.
- **Repeated Citations**: Redundant citations contravene the multi-hop requirement of sourcing from multiple documents.
- **Extreme Quotes**: Quotes that are either too terse (under five words) or excessively lengthy (spanning an entire document) lack utility.

Table 3 presents the substantial incidence rates

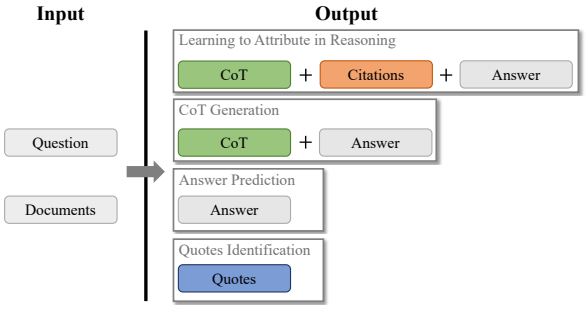

Figure 1: Comparison of the proposed auxiliary tasks.

of each error type, which could negatively impact fine-tuning effectiveness. After filtering, we obtain a training dataset of 1,358 samples, referred to as *MuSiQue-Attribute*. The statistics of the MuSiQue-Attribute training set are outlined in Table 4. It is important to note that the hop distribution in MuSiQue-Attribute is skewed. This skewness arises both because generated CoT for questions with more hops is more prone to errors and because such questions represent a smaller fraction of the original MuSiQue training set. Additionally, we conducted a human evaluation of the MuSiQue-Attribute quality in Appendix A.

3 Learning to Attribute in Reasoning

One intuitive approach to enhancing the multi-hop reasoning capabilities of LMs is to fine-tune them on our curated MuSiQue-Attribute, thereby teaching them to integrate attribution into their reasoning processes, specifically to generate CoC. Despite the simplicity of this method, our subsequent analysis in Section 4.4 demonstrates that this direct approach fails to produce robust results.

Multi-Task Learning. Beyond simply fine-tuning LMs on the MuSiQue-Attribute to learn to attribute in reasoning (denoted as **LA**), we propose three auxiliary tasks that serve as simplified analogs of LA. These tasks are designed to train LMs in conjunction with LA to enhance their proficiency in attribution-based reasoning:

- **Answer Prediction** (**AP** for short): This task focuses on direct answer prediction without the need for an explicit reasoning process. AP is intended to help LMs internalize the reasoning needed for straightforward questions where CoT is not required.
- **CoT Generation** (**CG** for short): In the CG task, models are trained to generate a CoT before providing an answer. This is aimed at developing LMs' abilities to reason explicitly and methodically across multiple pieces of information for complex questions.
- **Quotes Identification** (**QI** for short): This task trains models to pinpoint critical quotes for reasoning. QI is designed to fine-tune the ability of LMs to filter out irrelevant details and zero in on the pertinent segments of text, thereby sharpening the accuracy of reasoning.

Figure 1 illustrates the distinctions between our primary LA task and the three auxiliary tasks.

Data Augmentation. A recognized limitation of direct fine-tuning on our MuSiQue-Attribute is the potential for models to develop biases, such as favoring certain locations of relevant documents (Liu et al., 2023a), sensitive to a fixed number of documents, or accommodating only a narrow range of noise levels. To counteract these biases, we have devised the following data augmentation strategies:

- **Distractor Sampling**: By randomly selecting a varying number of irrelevant documents, we modify the positioning of relevant documents and the total document count within the context. This approach also mimics the fluctuating noise levels encountered in real-world scenarios, training language models to cope with noisy contexts effectively.
- **Document Shuffling**: Reordering the documents helps to remove any superficial positional cues that could lead to reasoning bias. For example, this ensures that models do not learn to associate the sequence of relevant documents with a fixed reasoning chain sequence.

These data augmentation strategies are applied in sequence for each training instance.

4 Experiments

4.1 Datasets

Our method's effectiveness in multi-hop reasoning is assessed on the following datasets: **HotpotQA** (Yang et al., 2018), **2WikiMultiHopQA** (**2Wiki** for short) (Ho et al., 2020), and **MuSiQue** (Trivedi et al., 2022). For each question, we provide a context composed of shuffled relevant and irrelevant documents. These irrelevant documents are the official retrieved distractor documents. We adopt the development and test sets from Trivedi et al. (2023) for evaluation, which contains 100 and 500 examples respectively. The results we present are the mean values from three separate trials, each with a distinct random seed.

	Model	MuSiQue				2Wiki					HotpotQA		
		Overall	2-Hop	3-Hop	4-Hop	Overall	Compositional	Inference	Comparison	Bridge-Comparison	Overall	Bridge	Comparison
5-Shot	ChatGPT												
	+ AO	15.8	16.1	14.3	17.4	46.2	19.0	68.1	30.5	71.4	51.0	49.0	60.2
	+ CoT	36.2	34.6	38.3	37.0	55.2	24.1	89.1	28.4	90.5	56.8	56.1	60.2
	+ CoC	37.0	37.0	37.0	37.0	55.4	27.8	89.1	28.4	88.6	58.6	57.5	63.6
	Claude-instant												
	+ AO	26.2	25.2	27.3	27.2	47.0	19.0	68.9	32.5	70.5	54.4	52.7	62.5
	+ CoT	26.0	27.2	27.9	19.6	40.8	20.3	80.7	15.7	58.1	20.2	23.5	4.5
	+ CoC	32.2	32.7	30.5	33.7	53.4	36.7	90.8	22.3	81.9	54.2	55.3	48.9
	LongChat												
	+ AO	6.7	7.0	4.1	10.1	26.8	12.0	3.4	48.7	47.3	32.3	34.0	24.6
	+ CoT	9.7	12.1	6.7	8.3	27.1	17.8	7.2	43.7	40.6	38.5	38.2	40.2
	+ CoC	11.0	13.3	8.7	8.7	24.5	19.0	5.9	42.9	27.9	39.1	38.9	39.8
	LongLoRA												
	+ AO	0.2	0.4	0.0	0.0	7.7	9.0	4.6	13.2	1.3	16.9	16.3	19.3
	+ CoT	0.0	0.0	0.0	0.0	15.1	5.9	0.8	28.9	27.3	11.4	11.0	13.3
	+ CoC	0.0	0.0	0.0	0.0	8.3	3.2	1.3	15.7	14.6	4.4	4.0	6.4
	Vicuna												
	+ AO	0.1	0.1	0.0	0.0	20.5	5.2	5.1	31.9	47.6	22.3	24.2	13.6
	+ CoT	0.0	0.0	0.0	0.0	27.7	14.7	7.6	48.5	43.5	30.5	32.0	23.1
	+ CoC	0.0	0.0	0.0	0.0	28.4	20.6	7.2	49.3	35.2	33.1	34.7	25.8
0-Shot	AttrLoRA												
	+ AO	32.9	35.7	27.1	34.8	**49.2**	48.7	28.7	54.9	59.0	51.5	51.0	54.2
	+ CoT	37.9	41.6	35.5	31.9	46.8	48.4	25.3	53.5	52.4	50.9	51.2	49.2
	+ CoC	**38.1**	42.7	35.7	29.7	47.4	46.0	27.4	57.7	53.3	**52.1**	52.6	49.6

Table 5: Exact-Match (EM) results on three multi-hop reasoning datasets. The best small-scale long-context LM results are in **bold** and the best baseline results are underlined.

To understand the broader impact of enhancing multi-hop reasoning on LMs' overall capabilities, we also conduct evaluations on general instruction-following benchmarks, namely **MT-Bench** (Zheng et al., 2023b) and **AlpacaEval** (Li et al., 2023c).

4.2 Models

The following long-context baselines are chosen in our experiments.

- **ChatGPT (Brown et al., 2020).** We choose `gpt-3.5-turbo-1106`, which supports a window size of 16K tokens.
- **Claude-instant (Bai et al., 2022).** We choose `claude-instant-1.2`, which has a window size of 100K tokens.
- **LongChat (Li et al., 2023a).** We use `longchat-7b-16k`, a 7B fine-tuned LLaMA model (Touvron et al., 2023a). It has a window size of 16K tokens.
- **LongLoRA (Chen et al., 2023b).** We use `LongAlpaca-7B-16k`, which has a window size of 16K.
- **Vicuna (Zheng et al., 2023b).** We use `vicuna-7b-v1.5-16k`, a 7B fine-tuned LLaMA-2 model (Touvron et al., 2023b). It supports a window size of 16K tokens.

We prompt all models with 5-shot to evaluate their multi-hop reasoning performance. These 5 demonstrations are randomly sampled from the 20 annotated training examples provided by Trivedi et al. (2023). If the input length exceeds the window size, we drop the last demonstration until the input length fits. The prompts we used are presented in Appendix D.

Our model **AttrLoRA** is fine-tuned on `vicuna-7b-v1.5-16k` with LoRA (Hu et al., 2022), following hyper-parameters used in FastChat (Zheng et al., 2023b). For its training data, we perform augmentations to double the training data for all tasks in Section 3 except for the QI task. Note that we subsample same-sized instruction-tuning data from the Alpaca dataset (Taori et al., 2023) and mix it with the reasoning data. These instruction-tuning data serve the purpose of minimizing the risk of hampering other abilities Vicuna already possesses before fine-tuning.

4.3 Main Results

Effectiveness and Robustness of Reasoning with Attributions. The results in Table 5 underscore the efficacy of our CoC prompting across three multi-hop reasoning datasets, benchmarked against five baselines. In 77% of the evaluated cases (disregarding instances of near-zero model performance) CoC outperforms CoT. Notably, Claude-instant exhibits strong results with AO, and its performance diminishes when CoT is used. However, CoC not only mitigates this decline but also attains results on par with AO, demonstrating the robustness of attribution-based reasoning. For a detailed analysis

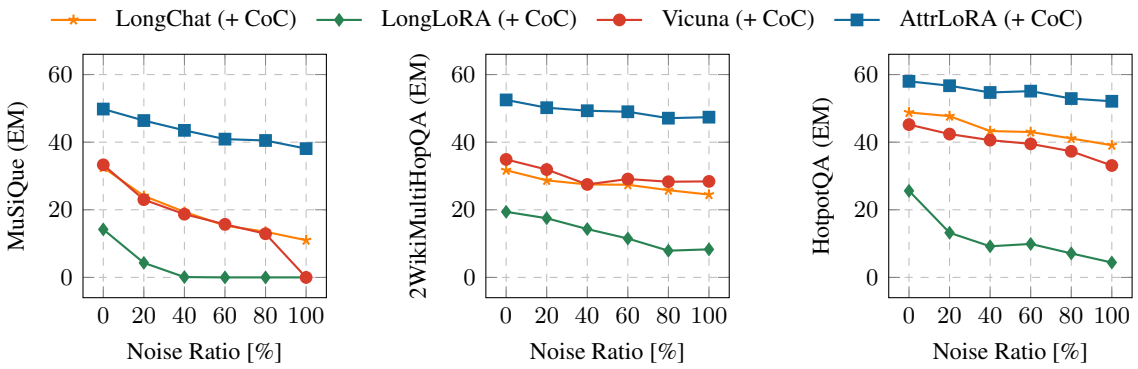

Figure 2: Exact-Match (EM) results of different models under various noise levels in three multi-hop reasoning datasets. Note that all models except our AttrLoRA use 5-shot prompting. A higher noise ratio indicates more distractors, i.e., irrelevant documents, are presented in the context of both the test instance and the demonstrations.

Model	MT-Bench (Score)	AlpacaEval (Win Rate)
ChatGPT	8.245	9.178%
Claude-instant	8.131	15.664%
Vicuna	6.068	5.415%
+ Alpaca Data	4.850	3.287%
AttrLoRA	4.978	3.106%

Table 6: Results on general instruction-following benchmarks. "+ Alpaca Data" is a Vicuna-7B model continued fine-tuning on Alpaca data.

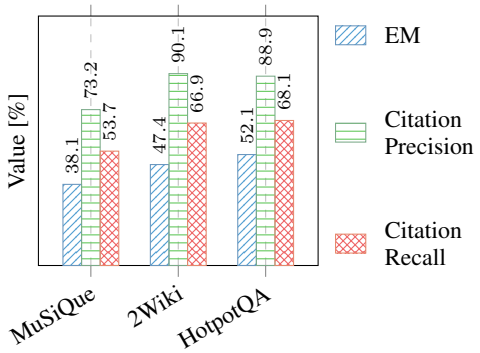

Figure 3: Multi-hop reasoning performance vs. citation precision and recall of AttrLoRA.

of CoC's robustness, particularly against varying degrees of contextual noise, see Appendix C.

Performance of AttrLoRA Against Proprietary Models. Table 5 presents a zero-shot performance comparison between our AttrLoRA and five-shot outcomes from various baselines. AttrLoRA surpasses baselines of comparable scale by an average margin of over 20 points. It exceeds the performance of two notable proprietary models on MuSiQue and delivers closely competitive results on the other two benchmarks.

In particular, AttrLoRA with AO achieves superior results on 2Wiki and surpasses CoT on HotpotQA. This can be attributed to the relative simplicity of these datasets, where explicit reasoning does not significantly enhance performance. For instance, CoT's advantage is noticeably smaller on these datasets compared to MuSiQue for both the ChatGPT and Claude-instant. Additionally, according to Jiang and Bansal (2019), over half of the "bridge-type" questions in HotpotQA contain shortcuts, which can locate the answer by keyword matching, circumventing the need for the intended two-hop reasoning. Similarly, 2Wiki's predictable nature, due to its question construction from a limited set of rules, simplifies the task for LMs. Another contributing factor is that AttrLoRA is trained on MuSiQue-Attribute, which does not encompass the full range of question types found in 2Wiki and HotpotQA, such as "comparison-type" questions.

Resilience of AttrLoRA to Noisy Contexts. A key aspect of AttrLoRA is its robustness to contextual noise. To investigate this, Figure 2 illustrates AttrLoRA's performance against varying degrees of synthesized noise. This synthesized noise is implemented by adding varied numbers of random irrelevant documents to the context. The data indicates that while the performance of baseline models markedly declines with increased noise, e.g., Vicuna drops by over 30 points on MuSiQue, AttrLoRA shows greater resilience, with a reduction of only about 10 points.

Impact of Attribution Learning on General Abilities. Our investigation extends beyond multi-hop reasoning to examine how attribution learning affects AttrLoRA's general instruction-following capabilities post-fine-tuning, as compared to the Vicuna baseline. The results in Table 6 from two

Model	MuSiQue	2Wiki	HotpotQA
Vicuna (5-Shot)	0.00	28.4	33.1
+ AP	32.3	47.8	52.1
+ CG	37.3	45.3	50.7
+ LA	37.3	46.9	52.1
+ QI	38.1	47.4	52.1

Table 7: Ablation study on multi-task learning.

Model	MuSiQue	2Wiki	HotpotQA
Vicuna (5-Shot)	0.00	28.4	33.1
+ AP	27.4	44.9	50.9
+ Augmentation	32.3	47.8	52.1
+ CG	28.1	37.7	49.4
+ Augmentation	30.5	37.3	50.4
+ LA	29.1	38.8	49.0
+ Augmentation	30.1	37.1	49.9

Table 8: Ablation study on data augmentation.

instruction-following benchmarks reveal that fine-tuning slightly compromises abilities beyond multi-hop reasoning in a 7B model due to capacity constraints. However, a closer analysis reveals that over 98% of the performance decrease is attributed to fine-tuning with Alpaca data ("+ Alpaca Data"), while multi-hop reasoning data incurs less than a 2% detriment. This is because the quality of Alpaca data is inferior to Vicuna's, with the former being single-turn GPT-3 synthesized and the latter comprising multi-turn human-bot conversations.

4.4 Analysis

Attribution Quality of AttrLoRA. Drawing from the insights of Gao et al. (2023), we scrutinize the citation precision and recall for AttrLoRA, as presented in Figure 3. The model demonstrates high precision, indicating its proficiency in correctly attributing statements to pertinent documents. Nonetheless, the moderate recall highlights that AttrLoRA does not consistently identify all relevant documents, a potential consequence of the disconnected reasoning patterns observed in MuSiQue-Attribute (detailed in Appendix A). Further exploration of the correlation between attribution quality and reasoning performance is in Appendix B.

The Effectiveness of Multi-Task Learning. Our ablation study in Table 7 assesses our multi-task learning approach. Results indicate a marked enhancement in Vicuna's reasoning capabilities upon fine-tuning with our dataset ("+ AP"). However, explicitly training Vicuna to generate CoT ("+ CG") yields mixed outcomes: It benefits performance on

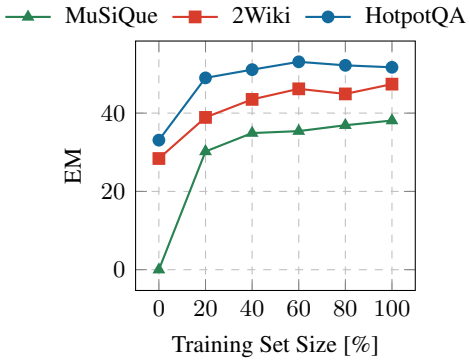

Figure 4: The impact of scaling fine-tuning data size.

MuSiQue but adversely affects results on 2Wiki and HotpotQA. This discrepancy can be attributed to the relative ease of the latter datasets, where simpler questions and shortcuts reduce the effectiveness of complex reasoning strategies, as discussed in Section 4.3. Importantly, integrating the LA task ("+ LA") mitigates the performance drops associated with CoT and notably boosts MuSiQue scores. This implies that attributions are instrumental in enabling the model to reason over complicated questions without compromising its ability to handle simpler queries. Finally, the addition of the QI task ("+ QI") appears to further refine the model's multi-hop reasoning proficiency, underscoring the value of our multi-task learning framework.

The Effectiveness of Data Augmentation. We explore the impact of our data augmentation strategy, intentionally omitting the QI task, as it alone is insufficient for training models to conduct multi-hop reasoning. The data in Table 8 demonstrates that including augmented data generally enhances model performance across various datasets. However, augmenting CG and LA data does not yield improvements on 2Wiki. In this case, the model readily learns from a limited amount of annotated data due to the simplicity of the automatically generated questions within 2Wiki. Conversely, on MuSiQue and HotpotQA, which feature more complex and varied human-crafted questions, the model benefits from exposure to a larger dataset to accommodate the diversity of question formulations.

The Effectiveness of Scaling Fine-Tuning Data. In Figure 4, we investigate how the expansion of fine-tuning data influences model performance. It is evident that incorporating additional data steadily enhances performance on MuSiQue and 2Wiki, while optimal results are attained with just 60% of our data for HotpotQA. This fact suggests that more

Question: Who is the mascot of the university related to Randy Conrads? **Document** [4](Title: Benny Beaver): Benny Beaver is the official mascot of Oregon State University and winner of the 2011 Capital One Mascot of the Year write - in campaign. . . . **Document** [7](Title: Randy Conrads): Randy Conrads attended Oregon State University, graduating in 1972 with a bachelor's degree in industrial engineering. . . . (Other irrelevant documents are omitted.) **Vicuna**: the university. ✘ **AttrLoRA**: Randy Conrads attended Oregon State University [7]. The mascot of Oregon State University is Benny Beaver [4]. The answer is: Benny Beaver ✔	**Question**: What is the record label of the co-writer and recording artist of Permission to Fly? **Document** [8](Title: No Ordinary Girl): No Ordinary Girl is the debut album by Jordan Pruitt, released in the United States on February 6, 2007 by Hollywood Records. The album debuted and peaked at number sixty-four . . . **Document** [19](Title: Permission to Fly): Permission to Fly is the second and final studio album by American singer-songwriter Jordan Pruitt. On July 22, 2008, the album was released . . . (Other irrelevant documents are omitted.) **Vicuna**: The answer is: The answer is a question. ✘ **AttrLoRA**: The co-writer and recording artist of "Permission to Fly" is Jordan Pruitt [19]. The record label of Jordan Pruitt is Hollywood Records [8]. The answer is: Hollywood Records ✔

Table 9: Two examples from MuSiQue test set where Vicuna fails and AttrLoRA succeeds. We manually annotate quotes in blue and citations in orange. Correct predictions are marked by ✔ and incorrect ones are by ✘.

complex question answering, involving additional reasoning steps like MuSiQue and 2Wiki, demands a larger dataset. An intriguing discovery is that using a mere 20% of our data achieves approximately 85% of the peak performance. This highlights the efficiency of fine-tuning: Even a modest subset of multi-hop reasoning examples can significantly boost the model's reasoning capabilities.

Case Study. Table 9 presents a comparative case study where Vicuna and AttrLoRA are both prompted to generate CoC. Within the provided examples, AttrLoRA successfully produces coherent CoT and precisely attributes each claim. In contrast, Vicuna yields answers without engaging in an explicit reasoning process.

5 Related Work

Multi-Hop Reasoning. Multi-hop reasoning in open-domain question answering requires the synthesis and analysis of disparate facts across various documents to formulate a response. Key datasets in this field include HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022), which predominantly adopt a reading comprehension framework with pre-retrieved documents supplied by the creators. Traditional approaches often utilize a selector-reader model (Zhang et al., 2023; Zhu et al., 2021), where the selector is tasked with pinpointing relevant documents from the provided set, and the reader constructs an answer based on these selections.

Recent advances, however, pivot towards a paradigm that leverages long-context LMs (Khot et al., 2023; Trivedi et al., 2023). In this approach, the role of the selector is phased out, and instead, the entirety of the retrieved documents is processed by a long-context LM, which acts as the reader. Our study aligns with this emergent research trend, particularly focusing on the use of attributions to enhance the performance of multi-hop reasoning within this long-context LM framework.

Context Utilization. The recent advent of long-context LMs has shown promise (Li et al., 2023a; Zheng et al., 2023b; Chen et al., 2023b). However, these models often struggle with noisy contexts. Shi et al. (2023) demonstrate that superfluous sentences can significantly disrupt mathematical reasoning. Liu et al. (2023a) identify a relevant document position bias in multi-document QA. Wu et al. (2024) show that LMs could be easily distracted by retrieved irrelevant inputs.

To mitigate the impact of irrelevant context, Shi et al. (2023) prompt models to disregard such information and adopt self-consistency techniques (Wang et al., 2023). Creswell et al. (2023) suggest a two-stage approach that focuses on fact selection prior to reasoning. Echoing this approach, Yu et al. (2023) introduce Chain-of-Note which entails reviewing document relevance before providing an answer. Meanwhile, Yoran et al. (2023) examine automatic data generation for training more robust models. Our research contributes to this domain by investigating the use of attributions as a novel method for effective context utilization.

Language Models Attribution. Attribution in language models constitutes a nascent area of study, primarily aimed at identifying and mitigating hallucination (Li et al., 2023b). A line of research concentrates on post-retrieval answering: Models provide responses based on retrieved results with cited attributions (Nakano et al., 2021; Menick et al., 2022; Gao et al., 2023). Our research emerges from this foundation but diverges in its applica-

tion; We focus on multi-hop reasoning rather than hallucination reduction. Moreover, we delve into optimizing training methodologies to maximize the efficacy of scarce attribution annotations.

6 Conclusion

This study demonstrates that long-context LMs face challenges with multi-hop reasoning within noisy contexts. We introduce a reasoning paradigm that incorporates attributions, which significantly improves the reasoning capabilities of long-context LMs. Alongside, we contribute a new dataset annotated with attributions and study training strategies tailored for multi-hop reasoning. Our comprehensive experiments across five models and five benchmarks validate the superiority of our approach in enhancing multi-hop reasoning performance.

Limitations

The proposed reasoning with attributions method currently leverages citations and quotations, leaving room for future exploration of other attribution forms, such as URLs. This work concentrates on refining training strategies, yet the potential of custom model architectures for this task remains untapped. Additionally, our approach relies on contexts pre-supplied by dataset creators. Emerging research suggests that language models integrated with search engines can achieve enhanced outcomes. A promising avenue for further research lies in developing models that not only manage noisy contexts more effectively, as demonstrated in our work, but also actively engage with search tools to improve information retrieval.

Ethics Statement

Our dataset builds upon MuSiQue, adhering to its original copyright provisions; We distribute our supplementary annotations under the CC BY 4.0 license. While our model-generated annotations could potentially include misinformation or harmful content, our manual quality review in Appendix A did not encounter such instances. Throughout the human annotation phase, we did not gather any demographic data or information that could reveal the identity of the annotators. All annotators provided informed consent for the use of their annotations exclusively for research purposes.

Acknowledgements

This work was supported by National Key R&D Program of China (Project No. 2022ZD0161200, 2022ZD0161201). This work is also supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 24200223) and Hong Kong Innovation and Technology Commission Project No. ITS/228/22FP. This work was also partially funded by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK.

References

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *CoRR*, abs/2307.11088.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *CoRR*, abs/2308.14508.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *CoRR*, abs/1604.06174.

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.

Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. *CoRR*, abs/2208.14271.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023a. How long can open-source llms truly promise on context length?

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

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023c. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts.

Tianyang Liu, Canwen Xu, and Julian J. McAuley. 2023b. Repobench: Benchmarking repository-level code auto-completion systems. *CoRR*, abs/2306.03091.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, H. Francis Song, Martin J. Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *CoRR*, abs/2203.11147.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

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

Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, and Yuxiong He. 2021. Zero-infinity: breaking the GPU memory wall for extreme scale deep learning. In *International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2021, St. Louis, Missouri, USA, November 14-19, 2021*, page 59. ACM.

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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang,

and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863, Online. Association for Computational Linguistics.

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.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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?

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *CoRR*, abs/2310.01558.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *CoRR*, abs/2311.09210.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, and Shen Huang. 2023. Beam retrieval: General end-to-end retrieval for multi-hop question answering.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023a. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4862–4876. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering.

Entry	Pearson	Spearman	Kendall
MuSiQue			
EM vs. P	0.887*	0.951*	0.826*
EM vs. R	0.484*	0.481*	0.411*
2Wiki			
EM vs. P	0.917*	0.896*	0.766*
EM vs. R	-0.172	-0.052	0.024
HotpotQA			
EM vs. P	0.865*	0.868*	0.741*
EM vs. R	-0.275	-0.038	0.023

Table 10: Correlation coefficients between citation precision (denoted as **P**) or recall (denoted as **R**) and the multi-hop reasoning performance (**EM**). * denotes statistically significant ($p < 0.05$).

Model	MuSiQue		2Wiki		HotpotQA	
	CoT	CoC	CoT	CoC	CoT	CoC
LongChat	22.3	**21.5**	**6.9**	7.2	9.7	**9.7**
LongLoRA	24.7	**14.2**	16.4	**11.5**	33.5	**21.2**
Vicuna	**32.9**	33.3	10.1	**7.5**	16.3	**12.1**
AttrLoRA	12.1	**11.7**	7.9	**5.4**	6.0	**5.9**

Table 11: The performance range of different models in three multi-hop reasoning benchmarks when the noise ratio of the context goes from 0% to 100% and models are prompted with CoT or CoC. The more robust results are highlighted in **bold**.

A Human Assessment of MuSiQue-Attribute

We recruit a student possessing expertise in NLP and holding a Master's degree in Computer Science to annotate 100 random samples from our MuSiQue-Attribute. The tool we used for annotations is shown in Figure 5. The annotation results reveal that 19% of the CoT annotations exhibited reasoning issues despite correct final answers, a problem referred to as *reasoning unfaithfulness* (Creswell and Shanahan, 2022). These erroneous CoTs fall into three categories: 5.26% with `Disordered Steps` where reasoning steps are improperly arranged, 78.95% with `Missing Steps` indicating omitted essential reasoning steps, also known as *disconnected reasoning* (Trivedi et al., 2020), and 10.53% with `Incorrect Steps` containing invalid reasoning steps. Our assessment also identified that 9% of questions in the dataset allowed for *shortcuts* (Jiang and Bansal, 2019), enabling models to find answers by keyword matching without reasoning.

The study further showed that within incorrect CoT annotations, 47.37% pertained to 2-hop questions, 36.84% to 3-hop, and 15.79% to 4-hop questions. Considering the distribution of hops in our dataset, the unfaithfulness issue affected 11.84% of 2-hop CoTs, 38.89% of 3-hop CoTs, and 50% of 4-hop CoTs, suggesting a decrease in reliability of LMs with the increase in reasoning steps required.

B Attribution Quality and Reasoning Performance

Figure 6 presents a visualization of the relationship between citation precision or recall and multi-hop reasoning performance. Our analysis identifies a notable positive correlation between citation precision and multi-hop reasoning capabilities. Table 10 provides statistical support for this observation, with correlation coefficients indicating a significant positive correlation. These results suggest the potential of using citation precision as a reference-free proxy for assessing multi-hop reasoning performance, which could be facilitated by Natural Language Inference (NLI) methods (Gao et al., 2023).

C Robustness to Noisy Context

We assessed the resilience of CoT and CoC against varying degrees of contextual noise by evaluating the performance variability of different models across multi-hop reasoning benchmarks. Table 11 indicates that in 84% of the cases, CoC exhibits a smaller performance range compared to CoT. Notably, even in instances where CoC demonstrates a greater range, the disparity with CoT remains marginal. These findings suggest that CoC is generally more robust to noisy contexts than CoT.

D Prompting Details

In this section, we highlight some details when prompting long-context LMs with various strategies, e.g., AO, CoT, CoC and CoQ.

- Each prompting strategy has its own instructions.
- For CoT, CoC and CoQ, we add "Think step-by-step." to the end of the question.
- For few-shot prompting, we put demonstrations to different turns as the dialogue history.
- Each demonstration is formatted according to Table 1.

The instruction for AO is (unique descriptions are underlined):

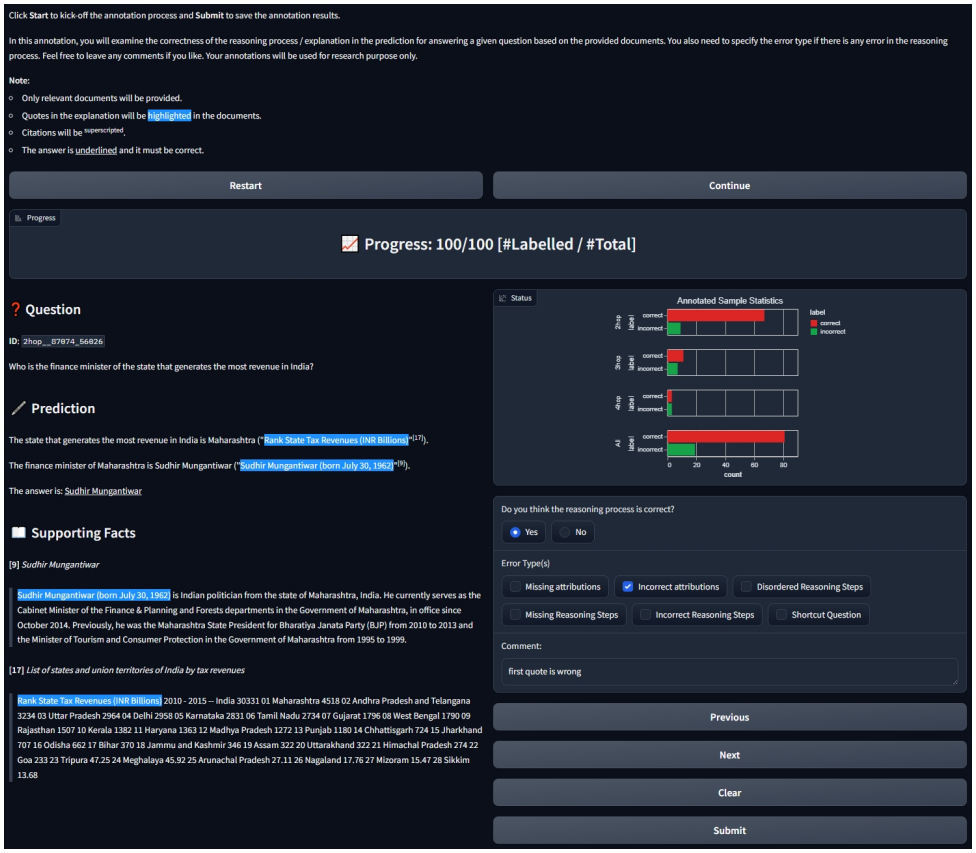

Figure 5: Screenshot of our human annotation tool.

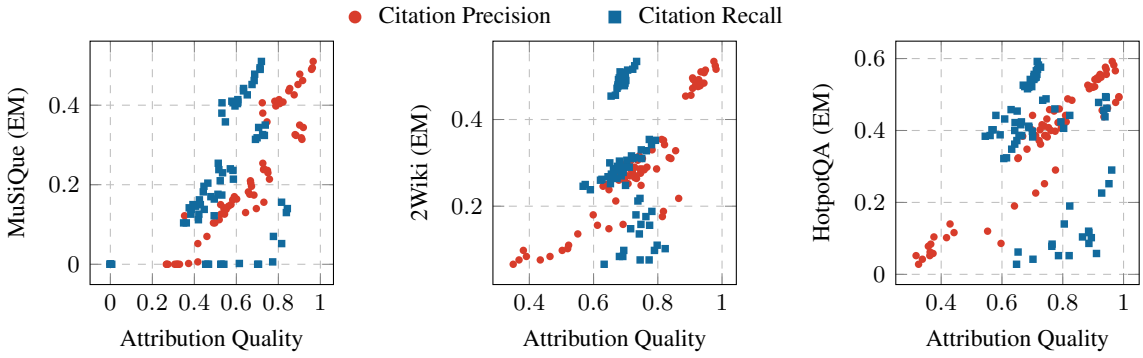

Figure 6: Attribution quality (citation precision and citation recall) vs. the multi-hop reasoning performance (EM). Data are collected from different trails of LongLoRA, LongChat, Vicuna and AttrLoRA.

> *Write an accurate and concise answer for the given question using only the provided search results (some of which might be irrelevant). Do not say anything other than the answer itself.*

The instruction for CoT is (differences to AO are underlined):

> *Write an accurate and concise answer for the given question using only the provided search results (some of which might be irrelevant). Start with an accurate, engaging, and concise explanation based only on the provided documents. Must end with "The answer is:". Use an unbiased and journalistic tone.*

The instruction for CoC is (differences to CoT are underlined):

> *Write an accurate and concise answer for the given question using only the provided search results (some of which might be irrelevant) and cite them properly. Start with an accurate, engaging,*

and concise explanation based only on the provided documents. Must end with "The answer is:". Use an unbiased and journalistic tone. Always cite for any factual claim.

The instruction for CoQ is (differences to CoC are underlined):

Write an accurate and concise answer for the given question using only the provided search results (some of which might be irrelevant) and cite them properly. Start with an accurate, engaging, and concise explanation based only on the provided documents. Must end with "The answer is:". Use an unbiased and journalistic tone. Always cite and extract word-for-word quotes for any factual claim.

E Filtering Implementation

The implementation of each filtering strategy is detailed as follows:

- **Incorrect Answer**: Leveraging the evaluation script from QuAC[2], we normalize both model predictions and official reference answers. We filter out data instances where the model prediction does not exactly match the answer.
- **Non-Existent Attributions**: Utilizing regular expressions, we extract all citations and verify if the predicted citations are present in the context. Each quote is checked for its exact presence in the cited document, and instances with fabricated citations or quotes are removed.
- **Incorrect Citations**: We assess the correctness of predicted citations using officially annotated supporting documents, ensuring the cited document is among the supporting documents. Any example with at least one incorrect citation is deleted.
- **Repeated Citations**: Regular expressions are used to extract all citations, and we check for duplicates. Examples with any duplicate citations are discarded.
- **Extreme Quotes**: We extract all quotes using regular expressions and tokenize them with NLTK[3]. Examples are kept only if all quotes contain more than five words but do not span entire cited documents.

F Training and Inference Cost

For fine-tuning using LoRA (Hu et al., 2022), we employed ZeRO-3 (Rajbhandari et al., 2021) optimization and gradient checkpointing (Chen et al., 2016) techniques. The model was trained on 8 NVIDIA A100 80GB GPUs for no more than 14 hours. Approximately 4M parameters were tuned, constituting 6.22% of the total parameters. For inference, each test set was processed in about 30 minutes using a single NVIDIA A100 80GB GPU.

[2] https://s3.amazonaws.com/my89public/quac/scorer.py

[3] https://www.nltk.org/