Chain-of-Thought Improves Text Generation with Citations in Large Language Models

Bin Ji, Huijun Liu*, Mingzhe Du, See-Kiong Ng

National University of Singapore
{jibin, mingzhe, seekiong}@nus.edu.sg, liuhuijun01@gmail.com

Abstract

Previous studies disclose that Large Language Models (LLMs) suffer from hallucinations when generating texts, bringing a novel and challenging research topic to the public, which centers on enabling LLMs to generate texts with citations. Existing work exposes two limitations when using LLMs to generate answers to questions with provided documents: unsatisfactory answer correctness and poor citation quality. To tackle the above issues, we investigate using Chain-of-Thought (CoT) to elicit LLMs' ability to synthesize correct answers from multiple documents, as well as properly cite these documents. Moreover, we propose a Citation Insurance Mechanism, which enables LLMs to detect and cite those missing citations. We conduct experiments on the ALCE benchmark with six open-source LLMs. Experimental results demonstrate that: (1) the CoT prompting strategy significantly improves the quality of text generation with citations; (2) the Citation Insurance Mechanism delivers impressive gains in citation quality at a low cost; (3) our best approach performs comparably as previous best ChatGPT-based baselines. Extensive analyses further validate the effectiveness of the proposed approach.

Introduction

Large Language Models (LLMs) have gained extensive research attention due to their powerful capability to understand human instructions and generate responses accordingly (Zhao et al. 2023). However, these responses may be unreliable since LLMs suffer from hallucinations, which puts forward demands to authenticate their factuality. Aimed at this issue, the research topic of enabling LLMs to generate texts with citations is brought to the public (Gao et al. 2023). It is easy for a human being to verify the factuality of LLMs' generations by checking their citations.

The majority of existing studies focus on adding citations to LLMs' generations in the scenarios of commercial search engines, *e.g.*, Bing Chat[1], and they use closed-source LLMs like ChatGPT[2] and GPT-4 (OpenAI 2023). Unfortunately, it is hard to automatically evaluate the generation correctness and the citation quality in these scenarios, which instead calls for low-efficiency and high-cost human evaluations (Liu, Zhang, and Liang 2023). Motivated by the above fact, Gao et al. (2023) propose the first reproducible benchmark, ALCE, to automatically evaluate LLMs' generations with citations. ALCE consists of three datasets and defines three automatic evaluation metrics, *i.e.*, Fluency, Correctness, and Citation Quality. Unlike those search engine-based studies, ALCE assumes natural-language questions paired with relevant retrieval corpora, which necessitates comprehensive systems to generate answers to questions using retrieved documents, as well as properly cite these documents. A series of LLMs, *e.g.*, ChatGPT, LLaMA (Touvron et al. 2023), Vicuna (Chiang et al. 2023), and Oasst (Köpf et al. 2023), have shown their exceptional abilities on ALCE. However, they still expose several limitations. One of the primary limitations lies in that they struggle to synthesize multiple documents in context and properly cite them (Gao et al. 2023),

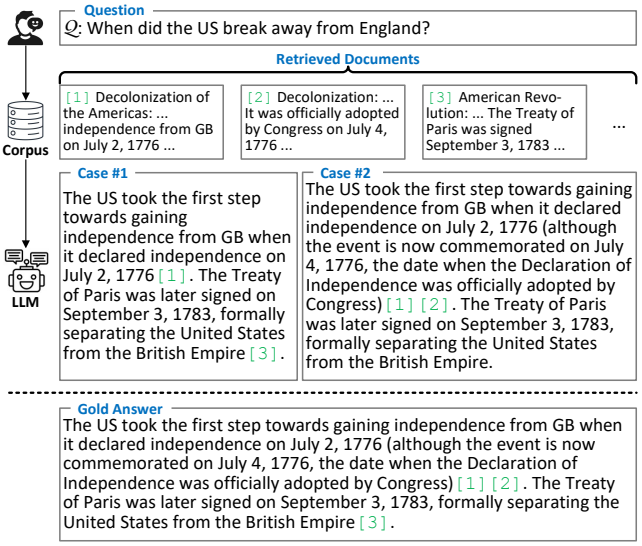

Figure 1: Case studies of LLMs' outputs. We re-use the ALCE example (Gao et al. 2023). **Case #1** shows that the LLM fails to consider the document [2] in its generated answer. **Case #2** exhibits that the LLM misses citing the document [3].

*Corresponding author

[1]https://www.bing.com/new

[2]https://chat.openai.com/chat

which in turn leads to unsatisfactory answer correctness and poor citation quality. We report two case studies in Figure 1.

We claim that the lack of generation process instruction is one cause of the unsatisfactory answer correctness and poor citation quality. Existing studies directly instruct LLMs to generate answers without providing them with the generation process like the Human Thought Process. For the ALCE benchmark, given a question and its retrieved documents, the Human Thought Process can be briefly formalized as (1) understand the question and the documents, and then find out relevant documents that can answer the question[3]; (2) for each relevant document, generate a statement accordingly to answer the question from some aspects and properly cite the document; (3) merge, rank, and concatenate these generated statements to obtain the final answer.

Intending to improve answer correctness and citation quality, we explore using Chain-of-Thought (CoT) (Wei et al. 2022; Wang et al. 2023b) prompting to boost LLMs' abilities in generating answers to questions with citations. To be specific, we propose CoT prompting strategies, achieved by instructing LLMs to mimic the Human Thought Process when generating answers. Moreover, we find that whether CoT prompts are employed or not, LLMs, including ChatGPT and GPT-4, are still prone to not cite any document in their generations, harming the citation quality. To tackle this issue, we propose Citation Insurance Mechanism (CIM) to detect those statements without any citation and add citations to them. We explore two different CIM implementations, *i.e.*, LLM-based CIM and IR-based CIM[4].

We evaluate our approach on six open-source LLMs with parameter scales ranging from 13 billion to 70 billion. For each LLM, we explore using CoT to enhance four prompt strategies. Experimental results on the ALCE benchmark demonstrate that: (1) the CoT prompting strategy significantly boosts the quality of text generation with citations, establishing strong and reproducible baselines for future study; (2) the proposed CIM improves citation quality at a low cost; (3) our approach based on LLaMA-2-70B-Chat achieves comparable performance as previous state-of-the-art ChatGPT-based baselines. Quantitative and qualitative analyses further validate the effectiveness of the CoT prompting strategy and the CIM. The main contributions are summarized as follows:

(1) By mimicking human behaviour, we make the first attempt to guide LLMs to generate texts with citations using Chain-of-Thought prompting strategies.

(2) We propose an efficient and low-cost Citation Insurance Mechanism to guarantee promising citation quality.

(3) Our approach establishes strong and reproducible baselines for future study.

Related Work

Text Generation with Citations. It is a new paradigm brought by the development of LLMs. Previous studies mainly focus on practical applications of LLM-enhanced commercial systems. For instance, Bing Chat[5] and perplexity.ai[6] respond to user questions in natural language with references to Web pages. However, it is hard to automatically evaluate the quality of these generations. To facilitate the automatic evaluation of text generation with citations, Gao et al. (2023) propose ALCE, the first reproducible benchmark for automatically evaluating LLMs' generation with citations. ALCE collects three datasets, *i.e.*, ASQA (Stelmakh et al. 2022), QAMPARI (Rubin et al. 2022), and ELI5 (Fan et al. 2019), and provides multiple baselines built upon closed- and open-source LLMs.

Chain-of-Thought (CoT). CoT prompting (an instruction or a few CoT demonstrations) is a gradient-free technique of eliciting a coherent series of intermediate reasoning steps for each query from LLMs. Investigations on various LLMs, such as GPT (Brown et al. 2020; Ouyang et al. 2022; OpenAI 2023) and PaLM (Chowdhery et al. 2022; Anil et al. 2023) series models, demonstrate that CoT prompting enhances performance across a range of arithmetic, commonsense, and symbolic reasoning tasks (Wu, Zhang, and Huang 2023; Chen et al. 2023b; Wang et al. 2023a).

Wei et al. (2022) initially propose the few-shot CoT, which requires the manual design of a few demonstrations to facilitate the generation of reasoning paths. In contrast, Kojima et al. (2022) propose the zero-shot CoT, which employs a single zero-shot prompt that elicits reasoning paths from LLMs. By simply adding "*Let's think step by step.*" after each query, zero-shot CoT demonstrates that LLMs are capable zero-shot reasoners without the need for any manually constructed CoT demonstrations. Built upon the CoT study of Wei et al. (2022), numerous studies have advanced the development of CoT through various strategies, such as Auto-CoT (Zhang et al. 2023), least-to-most prompting (Zhou et al. 2023), and self-consistency CoT (Wang et al. 2023b). These advancements have significantly bolstered the performance of CoT prompting in tackling intricate tasks.

Unlike existing CoT studies that center on tasks having explicit reasoning steps (*e.g.*, arithmetic and commonsense), the task of text generation with citations doesn't have such steps, which brings us great challenges when generalizing the CoT prompting.

Task Formalization

The ALCE benchmark (Gao et al. 2023) presents the first comprehensive formalization of text generation with citations. We briefly revisit the formalization here.

Given a question $\mathcal{Q}$ and a large-scale knowledge base $\mathcal{D}$ (*e.g.*, Wikipedia) that contains knowledge to answer $\mathcal{Q}$, an LLM-based system is required to generate an answer $\mathcal{A}$ to $\mathcal{Q}$ with $\mathcal{D}$, and $\mathcal{A}$ consists of several statements:

$$\mathcal{A} = s_1 s_2 s_3 ... s_n$$

where each statement s_i contains factual claims summarized from relevant document texts, *i.e.*, $\mathcal{T}_i = \{t_{i,1}, t_{i,2}, ...\}$,

[3]Some of the retrieved documents may be irrelevant to the question (Gao et al. 2023).

[4]IR is short for "Information Retrieval".

[5]https://www.bing.com/new

[6]https://www.perplexity.ai

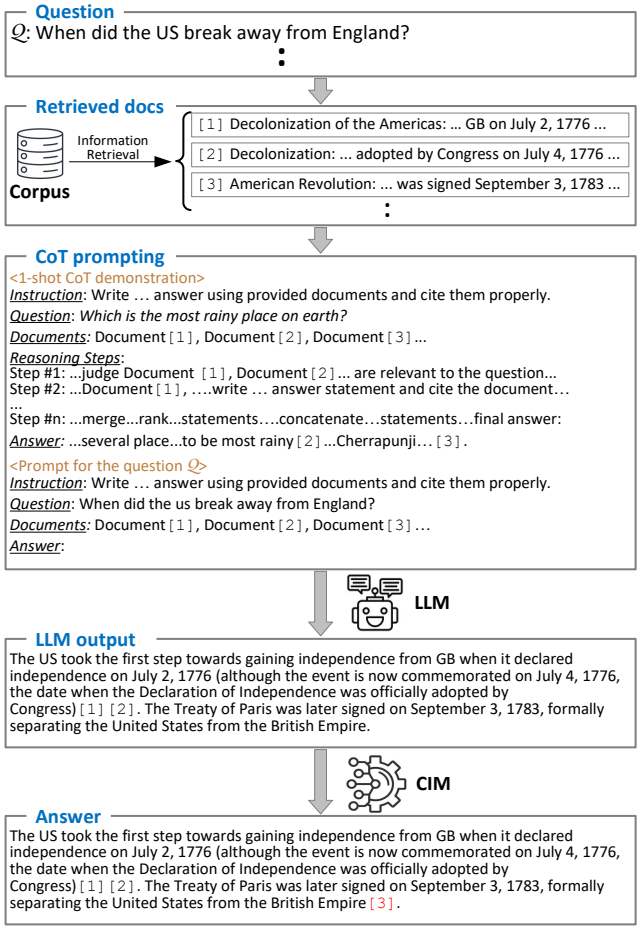

Figure 2: An overview of our approach. The "LLM output" is induced by the CoT prompting, which is fed to CIM to add the missing citation, *i.e.*, [3].

which are retrieved from $\mathcal{D}$ for $\mathcal{Q}$, and s_i needs to properly cite these documents[7]. ALCE regards each sentence of $\mathcal{A}$ as a statement.

Method

Figure 2 illustrates an overview of our approach. Given a question $\mathcal{Q}$, we first retrieve relevant documents using Information Retrieval (IR) methods; then we design a CoT prompting strategy to induce LLMs to answer $\mathcal{Q}$ by mimicking the Human Thought Process; at last, the proposed Citation Insurance Mechanism (CIM) detects those statements without any citation and adds relevant citations to them. The ALCE benchmark proves the effectiveness of IR methods like GTR (Ni et al. 2022) and BM25[8]. We follow this IR setting in our approach for fair comparisons.[9]

In the following sections, we first formalize our CoT prompting, then elaborate on the proposed CIM followed by extensive discussions.

CoT Prompting

CoT prompting mimics the Human Thought Process when solving a complicated reasoning task like the multi-step math word problem and the symbolic manipulation (Wei et al. 2022; Chen et al. 2023a,b). It decomposes the problem into several intermediate steps and solves each before giving the final answer.

Given a question $\mathcal{Q}$ and its retrieved documents $\mathcal{T} = \{t_1, t_2, ..., t_n\}$, we formalize the Human Thought Process of text generation with citations as follows:

(1) Understand $\mathcal{Q}$ and each document $t_i \in \mathcal{T}$ with the goal of finding out relevant documents that can answer the question. We use $\mathcal{T}'$ to denote the set of relevant documents, where $\mathcal{T}' \subseteq \mathcal{T}$.

(2) According to each document $t_j \in \mathcal{T}'$, generate a factual statement s_j to answer $\mathcal{Q}$ from some aspects and properly cite t_j using its unique document index, *i.e.*, [j]. This step generates a statement set $\mathcal{S} = \{s_1, s_2, ..., s_k\}$.

(3) Merge mutually entailed statements. Specifically, if two statements, *e.g.*, s_{k1} and s_{k2}, entails each other, *i.e.*, $s_{k1} \Leftrightarrow s_{k2}$, s_{k1} will be dropped but its citations will be merged to s_{k2}. We use $\mathcal{S}'$ to denote the statement set after merging.

(4) Rank and concatenate the statements included in $\mathcal{S}'$ to obtain the answer $\mathcal{A}$. This step aims to guarantee the fluency and coherence of $\mathcal{A}$.

We design 1-shot CoT demonstration based on the above Human Thought Process. As Figure 2 shows, the CoT demonstration consists of *Instruction*, *Question*, *Documents*, *Reasoning Steps*, and *Answer*, where the *Reasoning Steps* is the natural language description of the Human Thought Process. The CoT prompting strategy contains a 1-shot CoT demonstration and a prompt for the question $\mathcal{Q}$. We present a detailed CoT prompt in Appendix A.[10]

Although CoT has been comprehensively studied, extensive efforts are still necessary to generalize it to the task of text generation with citations. As far as we know, we are the first to generalize CoT to this task, where CoT prompts are built upon our formalization of the Human Thought Process. The formalization can also be used as the stepstone of future CoT prompt studies on this task.

Citation Insurance Mechanism

Whether CoT prompting strategies are employed or not, LLMs even ChatGPT and GPT-4 are still prone to not cite any document in their generated statements, which harms the citation quality. To tackle this issue, we propose the Citation Insurance Mechanism (CIM) to detect those statements without any citation and add relevant citations to them, as shown in Algorithm 1.

For these statements without any citation, we investigate two implementations of adding citations to them (the recite() function in Algorithm 1-L6), as shown below:

[7] ALCE requires that each s_i cites at least one and at most three citations. The citations should be in the format like [1] [2] [3].

[8] https://en.wikipedia.org/wiki/Okapi_BM25

[9] We use GTR for ASQA and QAMPRI, and BM25 for ELI5.

[10] The Appendix section can be found in the full paper at https://github.com/jibin5167/ALCE-CoT.

- **The LLM-based implementation.** We propose few-shot demonstration prompts to guide LLMs to add the missing citations. The LLM used to add citations is the same LLM used to generate answers. We present a real-world prompt example in Appendix A.
- **The IR-based implementation.** We propose to use GTR to search for the best matching document from the documents provided in the CoT prompt and cite it.

The main difference between the above two implementations lies in that: the LLM-based one can add more than one citation, but it may also fail to add any citation; in contrast, the IR-based one always adds one citation.

We conduct detailed analyses of the two implementations in the Analyses section.

Algorithm 1: Citation Insurance Mechanism

Input: $\mathcal{A}$
$\mathcal{A}'$ = ''
$\mathcal{X}$ = nltk.sent_tokenize($\mathcal{A}$) // Split $\mathcal{A}$ into statements
for $x \in \mathcal{X}$ **do**
  $\mathcal{C}$ = re.findall('\[\d+', x) // If x is a statement without any citation, then $\mathcal{C}$ is `Null`
  if $\mathcal{C}$ is `Null` **then**
    x' = recite(x) // Add relevant citations to x
    $\mathcal{A}' = \mathcal{A}' + x'$
  else
    $\mathcal{A}' = \mathcal{A}' + x$
  end
end
return $\mathcal{A}'$

Discussions

Can CoT prompts safely fit the limited context window? We claim that our CoT prompts can safely fit the context window on six open-source LLMs. LLMs have limited context window sizes, generally ranging from 2048 to 4096 tokens. To comply with the size limit, we adopt the 1-shot CoT prompt setting and only retrieve three documents for the 1-shot CoT demonstration and prompt for the question $\mathcal{Q}$.

Does CIM increase computation costs? It's clear that CIM increases computation costs, but we claim that the increased costs are actually very low. This is because CIM won't call computation-expensive LLMs or IR models until it detects statements without any citation. Take experiments on ASQA as examples, the ratios of statements without any citation range from 0.52% to 4.47% when using our CoT prompting strategy.

Does CIM always add correct citations? The answer is "No". We observe that both LLM-based and IR-based CIM may add incorrect citations. Here is a case study: when all the retrieved documents are irrelevant to a question $\mathcal{Q}$, LLMs tend to generate "*I cannot answer the question with the provided documents.*" as the answer to $\mathcal{Q}$. Under this condition, the LLM-based CIM is prone to cite all the provided documents, and the IR-based CIM cite one document. However, all these citations are incorrect.

Experiments

The ALCE Benchmark

ALCE (Gao et al. 2023) is the first reproducible benchmark for automatically evaluating LLMs' text generation with citations and allows for multiple citations for individual statements. It includes three datasets, *i.e.*, ASQA (Stelmakh et al. 2022), QAMPARI (Rubin et al. 2022), and ELI5 (Fan et al. 2019). It pre-defines three automatic evaluation metrics, *i.e.*, Fluency, Correctness (Correct.), and Citation Quality. We present more benchmark details in Appendix B.

Open-source LLMs

We evaluate our approach on six open-source LLMs (see Table 1) and report more LLM details in Appendix C.

Table 1: LLM details. We use LLaMA-2-13B to denote LLaMA-2-13B-Chat and LLaMA-2-70B to denote LLaMA-2-70B-Chat in other parts of this paper.

LLM	Window	LLM	Window
LLaMA-13B	2048	LLaMA-33B	2048
LLaMA-2-13B-Chat	4096	Oasst-33B	2048
Vicuna-13B	2048	LLaMA-2-70B-Chat	4096

CoT Prompting Strategy

In line with previous work (Gao et al. 2023), We explore applying CoT to four different prompting strategies:

(1) VANILLA. Provide the top-3 retrieved documents for each question.
(2) SUMM. Provide summaries instead of the full text of the top-10 retrieved documents for each question.
(3) SNIPPET. Provide snippets instead of full texts of the top-10 retrieved documents for each question.
(4) ORACLE. Provide three gold documents for each question.

The difference between SUMM and SNIPPET is that summaries are free-form, but snippets are spans extracted from documents. More details can be found in Appendix D.

Implementation Details and Baselines

We use four NVIDIA A100 40GB GPUs to evaluate our approach. Specifically, we use one GPU to run 13B LLMs, two GPUs to run 33B LLMs, and four GPUs to run the 70B LLM. For all experiments, We set the seed to 42, which is the default setting of ALCE. For each prompting strategy, we evaluate our approach on the six LLMs by setting the temperature value to 0.001, 0.1, 0.3, 0.5, 0.7, 0.9, and 1, respectively. For each type of experiment, we average the results of different temperature settings and report the averaged performance.

The LLM-based prompting approaches proposed by Gao et al. (2023) are the current state-of-the-art approaches. These approaches are evaluated with ChatGPT, LLaMA-13B, LLaMA-33B, Vicuna-13B, and Oasst-33B. We adopt these approaches as baselines and take additional two LLaMA-2 models into consideration, as shown in Table 1.

Table 2: Performance comparisons of baselines (columns marked by "base") and the proposed approach (columns marked by "ours") on ASQA.

LLM	Fluency		Correct.		Citation			
					Recall		Precision	
	base	ours	base	ours	base	ours	base	ours
VANILLA								
LLaMA-13B	68.4	**76.8**	26.9	**30.7**	10.6	**17.9**	**26.9**	25.4
LLaMA-2-13B	73.4	**82.0**	32.6	**37.4**	60.0	**66.5**	52.1	**59.2**
Vicuna-13B	**82.6**	79.8	31.9	**36.4**	51.1	**56.6**	50.1	**56.4**
LLaMA-33B	83.7	**84.2**	31.0	**35.1**	19.5	**24.4**	23.0	**27.9**
Oasst-33B	**82.9**	81.7	34.8	**35.9**	36.2	**41.4**	38.3	**46.8**
LLaMA-2-70B	**84.6**	82.7	39.4	**43.9**	62.7	**70.2**	58.4	**69.8**
SUMM								
LLaMA-13B	76.8	**82.3**	33.3	**35.1**	19.6	**27.4**	23.7	**30.3**
LLaMA-2-13B	**81.5**	79.4	42.9	**43.1**	58.7	**62.3**	50.4	**55.1**
Vicuna-13B	67.7	**81.2**	43.2	43.2	52.7	**56.9**	50.0	**56.8**
LLaMA-33B	72.0	**79.6**	33.1	**35.7**	34.7	**42.2**	35.2	**40.8**
Oasst-33B	74.3	**77.2**	**40.9**	39.4	45.5	**47.6**	44.0	**49.7**
LLaMA-2-70B	84.3	**89.6**	43.7	**44.6**	64.2	**71.4**	57.7	**68.0**
SNIPPET								
LLaMA-13B	72.0	**79.0**	31.3	**32.5**	18.2	**22.9**	21.1	**25.4**
LLaMA-2-13B	81.4	**83.1**	41.3	**43.1**	57.4	**61.2**	52.1	**55.7**
Vicuna-13B	**81.4**	77.2	42.1	**42.9**	53.4	**55.6**	48.7	**52.3**
LLaMA-33B	70.8	**78.4**	30.9	**33.7**	31.4	**39.1**	31.5	**37.2**
Oasst-33B	79.3	**80.2**	**40.1**	38.9	45.0	**49.6**	43.3	**44.5**
LLaMA-2-70B	82.7	**86.3**	**43.2**	42.9	64.2	**70.4**	60.2	**69.4**
ORACLE								
LLaMA-13B	69.5	**78.2**	34.3	**34.7**	10.8	**15.2**	15.8	**17.1**
LLaMA-2-13B	73.2	**77.4**	41.4	**43.0**	54.5	**58.2**	52.9	**56.2**
Vicuna-13B	72.9	**79.3**	42.5	**42.7**	52.2	**56.2**	50.7	**54.8**
LLaMA-33B	**82.6**	79.4	39.3	**40.1**	20.2	**27.7**	23.9	**31.2**
Oasst-33B	**85.4**	82.6	**44.3**	43.1	37.0	**44.8**	39.6	**45.4**
LLaMA-2-70B	**89.5**	86.2	44.1	**46.2**	67.4	**73.5**	69.8	**72.9**

Table 3: Performance comparisons of baselines (columns marked by "base") and the proposed approach (columns marked by "ours") on QAMPARI.

LLM	Correct.				Citation			
	Recall-5		Precision		Recall		Precision	
	base	ours	base	ours	base	ours	base	ours
VANILLA								
LLaMA-13B	9.7	**12.3**	9.1	**14.6**	6.7	**11.1**	7.1	**12.7**
LLaMA-2-13B	13.4	**16.8**	15.2	**18.3**	12.2	**14.4**	12.5	**17.7**
Vicuna-13B	14.0	**18.1**	15.9	**17.4**	**12.5**	11.6	13.4	**15.4**
LLaMA-33B	14.7	**21.3**	12.0	**15.2**	7.9	**11.1**	8.3	**12.3**
Oasst-33B	**15.5**	14.4	14.9	**16.8**	9.0	**12.2**	10.1	**14.1**
LLaMA-2-70B	19.2	**23.2**	18.2	**21.7**	17.8	**22.3**	19.1	**24.0**
SUMM								
LLaMA-13B	14.8	**19.2**	12.6	**17.9**	7.4	**15.1**	8.0	**14.2**
LLaMA-2-13B	19.2	**23.4**	17.5	**22.1**	15.3	**18.5**	14.2	**20.4**
Vicuna-13B	21.1	**25.6**	17.1	**19.8**	15.7	**17.2**	17.8	**19.3**
LLaMA-33B	19.0	**25.7**	14.8	**17.9**	**12.5**	12.4	15.0	**19.7**
Oasst-33B	21.0	**24.4**	17.5	**22.9**	12.9	**16.8**	16.6	**19.1**
LLaMA-2-70B	22.0	**23.6**	19.8	**22.0**	20.2	**23.1**	21.4	**23.7**
SNIPPET								
LLaMA-13B	17.7	**21.1**	15.7	**19.8**	8.8	**14.4**	9.9	**14.7**
LLaMA-2-13B	22.3	**26.4**	19.2	**22.8**	14.3	**17.9**	20.1	**24.3**
Vicuna-13B	21.9	**25.6**	18.2	**21.3**	16.8	**19.4**	19.7	**24.4**
LLaMA-33B	19.6	**24.4**	15.7	**22.7**	12.8	**18.4**	15.2	**16.7**
Oasst-33B	22.0	**24.2**	17.4	**23.8**	13.6	**18.4**	17.7	**19.8**
LLaMA-2-70B	22.6	**22.9**	20.4	**23.1**	19.8	**22.3**	21.2	**23.8**
ORACLE								
LLaMA-13B	16.8	**17.4**	15.4	**17.2**	7.7	**11.9**	8.3	**14.6**
LLaMA-2-13B	26.4	**30.1**	30.2	**33.5**	17.3	**21.5**	19.1	**21.6**
Vicuna-13B	25.9	**27.8**	28.4	**31.2**	15.8	**20.2**	16.8	**19.7**
LLaMA-33B	23.9	**26.0**	20.3	**22.7**	9.8	**15.4**	10.4	**17.2**
Oasst-33B	26.9	**27.4**	**26.0**	25.6	11.7	**15.1**	12.9	**16.3**
LLaMA-2-70B	32.2	**36.1**	31.9	**36.7**	22.6	**24.4**	21.4	**24.4**

Main Results

We report performance comparisons of our approach and baselines in Table 2, Table 3, and Table 4. For the baselines, we evaluate their results on the two LLaMA-2 models by ourselves and directly take the results of other LLMs reported by Gao et al. (2023) to facilitate fair comparisons. We have the following observations:

(1) On ASQA, our approach consistently outperforms the baselines in terms of Citation Recall, and it beats them regarding Correct. and Citation Precision in 43 out of 48 instances, as well as exhibits great advantages regarding Fluency in 16 out of 24 instances.

(2) On QAMPARI, our approach delivers consistent Citation Precision gains, and it beats the baselines in 68 out of 72 instances regarding the other three metrics, *i.e.*, Recall-5, Precision, and Citation Recall.

(3) On ELI5, our approach showcases improvements in terms of Correct. and Citation Precision, and performs better than the baselines in 42 out of 48 instances when comparing them from Fluency and Citation Recall perspectives.

These promising performance gains demonstrate the effectiveness of our CoT prompting strategy and the Citation Insurance Mechanism.

We also compare our best model with the ChatGPT-based baseline, which achieves the current state-of-the-art results. The comparison results are reported in Table 5, from which we can observe that our best approach achieves comparable Correct. and Citation Quality to current SOTA results, revealing that our approach can serve as a strong reproducible baseline for future study. Additionally, it further verifies the effectiveness of the CoT prompting strategy and the CIM.

Analyses

Performance against CoT Descriptions

This analysis aims to investigate if different natural language descriptions of the Human Thought Process affect performance. To this end, we manually implement five versions of the Human Thought Process and investigate their effectiveness, as shown in Figure 3. We can see that these descriptions induce quite different performances in all four evaluation metrics, indicating that LLMs are sensitive to prompt descriptions, which highlights the importance of well-designed and LLM-adapted prompts, the instability of manual prompts, as well as prompt engineering.

Table 4: Performance comparisons of baselines (columns marked by "base") and the proposed approach (columns marked by "ours") on ELI5.

LLM	Fluency		Correct.		Citation			
					Recall		Precision	
	base	ours	base	ours	base	ours	base	ours
VANILLA								
LLaMA-13B	50.0	**59.9**	3.9	**12.3**	3.1	**9.7**	5.3	**9.2**
LLaMA-2-13B	57.9	**62.1**	12.1	**14.5**	16.4	**22.8**	19.7	**26.3**
Vicuna-13B	58.2	**69.4**	10.0	**13.3**	15.6	**21.3**	19.6	**24.1**
LLaMA-33B	58.8	**71.2**	6.2	**8.8**	9.3	**15.1**	12.1	**16.4**
Oasst-33B	46.8	**59.0**	9.5	**12.1**	**16.0**	15.2	21.6	**25.6**
LLaMA-2-70B	58.2	**62.4**	12.4	**13.9**	46.5	**51.1**	44.4	**49.1**
SUMM								
LLaMA-13B	28.6	**42.1**	2.9	**8.4**	2.5	**7.1**	3.8	**6.8**
LLaMA-2-13B	26.4	**31.4**	6.1	**11.2**	9.9	**14.2**	14.3	**17.4**
Vicuna-13B	22.4	**37.2**	4.9	**10.2**	9.7	**11.7**	12.2	**15.8**
LLaMA-33B	23.3	**37.1**	3.0	**9.0**	6.2	**9.8**	8.2	**14.1**
Oasst-33B	24.8	**34.1**	3.9	**7.6**	12.3	**18.9**	16.3	**22.7**
LLaMA-2-70B	36.8	**42.4**	9.8	**12.7**	44.6	**49.7**	44.7	**48.4**
SNIPPET								
LLaMA-13B	48.4	**57.1**	5.7	**9.2**	5.8	**7.7**	7.6	**12.1**
LLaMA-2-13B	52.1	**57.3**	11.9	**14.5**	29.4	**34.5**	28.6	**34.6**
Vicuna-13B	48.1	**55.2**	11.2	**13.6**	27.2	**29.3**	27.9	**33.6**
LLaMA-33B	**53.2**	49.8	7.4	**12.1**	13.7	**17.4**	15.1	**18.1**
Oasst-33B	50.7	**55.4**	10.7	**14.8**	25.8	**30.1**	26.7	**29.9**
LLaMA-2-70B	55.4	**58.9**	13.4	**13.9**	44.7	**49.6**	42.1	**45.2**
ORACLE								
LLaMA-13B	**49.5**	47.5	6.4	**8.3**	3.7	**9.4**	6.5	**9.1**
LLaMA-2-13B	**47.4**	45.8	16.9	**22.7**	21.4	**24.2**	27.3	**33.1**
Vicuna-13B	41.6	**48.4**	17.1	**22.1**	20.2	**24.3**	26.5	**31.1**
LLaMA-33B	**63.7**	62.1	11.4	**15.1**	11.9	**17.2**	15.4	**18.9**
Oasst-33B	50.7	**54.2**	15.8	**18.8**	20.8	**26.3**	28.0	**31.4**
LLaMA-2-70B	**58.4**	56.1	17.6	**20.8**	52.4	**56.7**	52.1	**55.3**

Performance against CIM Implementations

This analysis aims to compare the LLM-based CIM and IR-based CIM. We solely compare the Citation Quality since CIM is designed to advance it. We conduct the investigation using the LLaMA-2-13B VANILLA setting and report the results in Figure 4. We observe that LLM-based CIM achieves better Citation Quality on ASQA and ELI5, as well as higher Citation Precision on QAMPARI. We attribute this to the fact that LLM-based CIM may add more than one citation to the statements without any citation, and it refuses to add any citation if it cannot find any document that supports the statement, which guarantees better Citation Quality. In contrast, the IR-based CIM always add one citation to the statement even if there are multiple documents that support the statement or no document that supports the statement.

Based on the above analysis, we use the LLM-based CIM in all the other experiments.

Performance against Temperature

This analysis investigates if the temperature affects the quality of LLMs' outputs. To this end, we set the temperature to 0.001, 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, and 1, respectively. We conduct investigations on ASQA using the LLaMA-2-13B VANILLA setting and report the results in Figure 5, which indicates that the temperature value has obvious effects on all four evaluation metrics, especially the Fluency. We attribute this to the fact that various temperature values enable LLMs to generate texts with various randomness, consequently affecting the fluency of LLMs' generations. As a result, the correctness and the citation quality follow the variations.

Moreover, we observe that the Citation Recall and Citation Precision exhibit similar variation trends, revealing tight correlations between the two evaluation metrics.

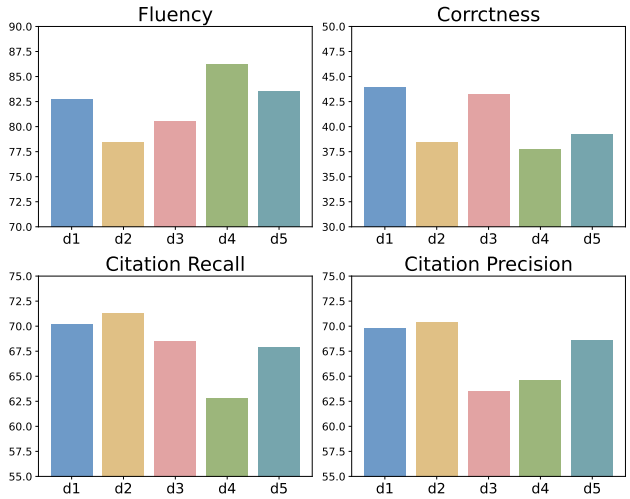

Figure 3: Performance comparisons of five different descriptions of the Human Thought Process, marked by d1, d2, d3, d4, and d5, respectively. We evaluate the performance on ASQA using the LLaMA-2-70B VANILLA setting.

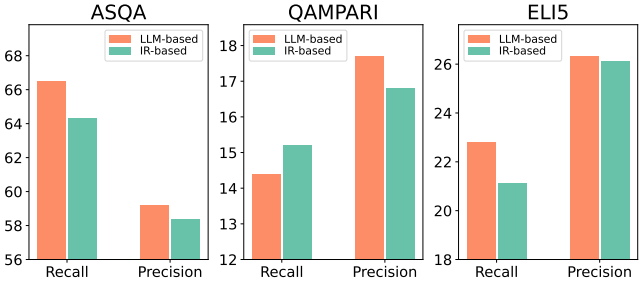

Figure 4: Comparisons of Citation Quality between LLM-based and IR-based CIM. We evaluate them on all datasets using the LLaMA-2-13B VANILLA setting.

Ablation Study

We ablate the CoT prompting strategy and the LLM-based CIM individually to explore their utilities. We report the ablation settings and results in Table 6, where:

- "`-CoT`" denotes removing the CoT prompting strategy. We use 1-shot in-context learning as a proxy.
- "`-CIM`" denotes removing the CIM. Our approach will do nothing to those statements without citations.
- "`base`" denotes conducting the above two ablations simutaneously.

Table 5: Comparisons with ChatGPT-based results.

	Strategy	Fluency	Correct.	Citation	
				Recall	Precision
ASQA	**ChatGPT**				
	VANILLA	66.6	40.4	73.6	72.5
	SUMM	70.0	43.3	68.9	61.8
	SNIPPET	69.8	41.4	65.3	57.4
	ORACLE	64.4	**48.9**	**74.5**	72.7
	LLaMA-2-70B (ours)				
	ORACLE	**86.2**	46.2	73.5	**72.9**

	Strategy	Correct.		Citation	
		Recall-5	Precision	Recall	Precision
QAMPARI	**ChatGPT**				
	VANILLA	20.8	20.8	20.5	20.9
	SUMM	23.6	21.2	23.6	25.7
	SNIPPET	24.5	21.5	22.9	24.9
	ORACLE	**37.0**	**36.9**	24.1	**24.6**
	LLaMA-2-70B (ours)				
	ORACLE	36.1	36.7	**24.4**	24.4

	Strategy	Fluency	Correct.	Citation	
				Recall	Precision
ELI5	**ChatGPT**				
	VANILLA	57.2	12.0	51.1	50.0
	SUMM	40.3	12.5	51.8	48.2
	SNIPPET	62.9	14.3	50.4	45.0
	ORACLE	**59.4**	**21.3**	**57.8**	**56.0**
	LLaMA-2-70B (ours)				
	ORACLE	56.1	20.8	56.7	55.3

We conduct ablation studies on three LLaMA models with various parameters. From Table 6 we observe that:

(1) CoT prompting strategy delivers significant performance gains across all evaluation metrics and all LLMs. We attribute it to the fact that CoT prompts guide LLMs to mimic the Human Thought Process, which has been proven to be effective in the Experiments section.

(2) CIM delivers more improvements in Citation Quality on relatively smaller LLMs. For example, it brings +2.4 Recall and +2.2 Precision gains on LLaMA-2-13B, while only +1.1 Recall and +1.9 Precision gains on LLaMA-2-70B. We attribute it to the fact that the larger the LLM's parameter scale, the more powerful the LLM's ability to properly add citations. Thus, fewer statements suffer from the citation missing problem in larger LLMs.

In addition, we note that the CIM also affects Fluency and Correctness. However, the CIM is designed to add missing citations, which does not affect the above two evaluation metrics theoretically. After checking LLMs' generations, we find that the LLM-based CIM may complete statements that are detected to have no citations and are not ended normally, consequently affecting the Fluency and Correctness.

We present a real-world case for better understanding: the LLM-based CIM completes "*Therefore, the answer is that*" and adds citations to it and outputs "*Therefore, the answer is that there are 34 state parks in Virginia*`[1][2][3]`."

Conclusion

In this paper, we investigate Chain-of-Thought (CoT) prompting strategies to improve LLMs' ability to generate text with citations. We formalize the Human Thought Process and carefully design CoT prompts accordingly. Moreover, we propose a Citation Insurance Mechanism (CIM) to detect those missing citations and properly cite them. Experimental results on six open-source LLMs and the ALCE benchmark demonstrate the effectiveness of our novel approach. And our approach induces LLaMA-2-70B to perform comparably to ChatGPT on the ALCE benchmark. Qualitative and quantitative analyses further prove the effectiveness of our approach.

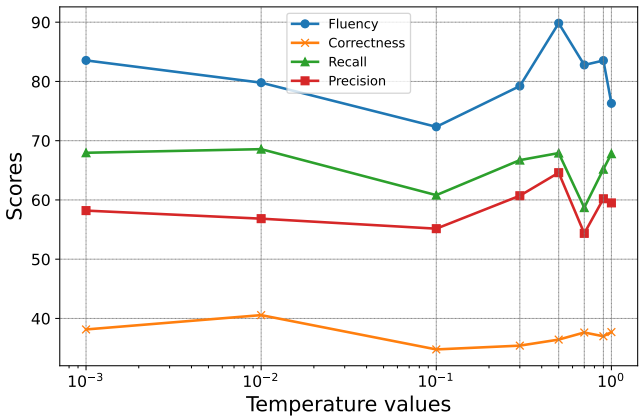

Figure 5: Performance comparisons of various temperature values, which are set to 0.001, 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, and 1, respectively. The x-coordinate is on a log scale.

Table 6: Ablation results on ASQA with various LLM settings.

Setting	Fluency	Correct.	Citation	
			Recall	Precision
LLaMA-2-13B	82.0	37.4	66.5	59.2
`-CoT`	76.3	33.2	62.8	54.4
`-CIM`	80.4	36.8	64.1	57.0
`base`	74.6	32.8	59.6	53.4
LLaMA-33B	84.2	35.1	24.4	27.9
`-CoT`	79.4	32.4	20.2	24.0
`-CIM`	81.9	34.9	23.3	26.4
`base`	81.4	30.8	19.9	23.4
LLaMA-2-70B	82.7	43.9	70.2	69.8
`-CoT`	77.8	39.6	63.3	58.9
`-CIM`	82.1	42.6	69.1	67.9
`base`	85.2	38.8	63.7	58.1

Acknowledgments

This research is supported by A*STAR, CISCO Systems (USA) Pte. Ltd and National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002).

Additionally, this research/project is supported by the National Research Foundation, Singapore under its Industry

Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; Chu, E.; Clark, J. H.; Shafey, L. E.; Huang, Y.; Meier-Hellstern, K.; Mishra, G.; Moreira, E.; Omernick, M.; Robinson, K.; Ruder, S.; Tay, Y.; Xiao, K.; Xu, Y.; Zhang, Y.; Abrego, G. H.; Ahn, J.; Austin, J.; Barham, P.; Botha, J.; Bradbury, J.; Brahma, S.; Brooks, K.; Catasta, M.; Cheng, Y.; Cherry, C.; Choquette-Choo, C. A.; Chowdhery, A.; Crepy, C.; Dave, S.; Dehghani, M.; Dev, S.; Devlin, J.; Díaz, M.; Du, N.; Dyer, E.; Feinberg, V.; Feng, F.; Fienber, V.; Freitag, M.; Garcia, X.; Gehrmann, S.; Gonzalez, L.; Gur-Ari, G.; Hand, S.; Hashemi, H.; Hou, L.; Howland, J.; Hu, A.; Hui, J.; Hurwitz, J.; Isard, M.; Ittycheriah, A.; Jagielski, M.; Jia, W.; Kenealy, K.; Krikun, M.; Kudugunta, S.; Lan, C.; Lee, K.; Lee, B.; Li, E.; Li, M.; Li, W.; Li, Y.; Li, J.; Lim, H.; Lin, H.; Liu, Z.; Liu, F.; Maggioni, M.; Mahendru, A.; Maynez, J.; Misra, V.; Moussalem, M.; Nado, Z.; Nham, J.; Ni, E.; Nystrom, A.; Parrish, A.; Pellat, M.; Polacek, M.; Polozov, A.; Pope, R.; Qiao, S.; Reif, E.; Richter, B.; Riley, P.; Ros, A. C.; Roy, A.; Saeta, B.; Samuel, R.; Shelby, R.; Slone, A.; Smilkov, D.; So, D. R.; Sohn, D.; Tokumine, S.; Valter, D.; Vasudevan, V.; Vodrahalli, K.; Wang, X.; Wang, P.; Wang, Z.; Wang, T.; Wieting, J.; Wu, Y.; Xu, K.; Xu, Y.; Xue, L.; Yin, P.; Yu, J.; Zhang, Q.; Zheng, S.; Zheng, C.; Zhou, W.; Zhou, D.; Petrov, S.; and Wu, Y. 2023. PaLM 2 Technical Report. arXiv:2305.10403.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Chen, J.; Chen, L.; Huang, H.; and Zhou, T. 2023a. When do you need Chain-of-Thought Prompting for ChatGPT? *arXiv preprint arXiv:2304.03262*.

Chen, Z.; Zhou, K.; Zhang, B.; Gong, Z.; Zhao, X.; and Wen, J.-R. 2023b. ChatCoT: Tool-Augmented Chain-of-Thought Reasoning on Chat-based Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14777–14790. Singapore: Association for Computational Linguistics.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; Schuh, P.; Shi, K.; Tsvyashchenko, S.; Maynez, J.; Rao, A.; Barnes, P.; Tay, Y.; Shazeer, N.; Prabhakaran, V.; Reif, E.; Du, N.; Hutchinson, B.; Pope, R.; Bradbury, J.; Austin, J.; Isard, M.; Gur-Ari, G.; Yin, P.; Duke, T.; Levskaya, A.; Ghemawat, S.; Dev, S.; Michalewski, H.; Garcia, X.; Misra, V.; Robinson, K.; Fedus, L.; Zhou, D.; Ippolito, D.; Luan, D.; Lim, H.; Zoph, B.; Spiridonov, A.; Sepassi, R.; Dohan, D.; Agrawal, S.; Omernick, M.; Dai, A. M.; Pillai, T. S.; Pellat, M.; Lewkowycz, A.; Moreira, E.; Child, R.; Polozov, O.; Lee, K.; Zhou, Z.; Wang, X.; Saeta, B.; Diaz, M.; Firat, O.; Catasta, M.; Wei, J.; Meier-Hellstern, K.; Eck, D.; Dean, J.; Petrov, S.; and Fiedel, N. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311.

Fan, A.; Jernite, Y.; Perez, E.; Grangier, D.; Weston, J.; and Auli, M. 2019. ELI5: Long Form Question Answering. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3558–3567. Florence, Italy: Association for Computational Linguistics.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023. Enabling Large Language Models to Generate Text with Citations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488. Singapore: Association for Computational Linguistics.

Honovich, O.; Aharoni, R.; Herzig, J.; Taitelbaum, H.; Kukliansy, D.; Cohen, V.; Scialom, T.; Szpektor, I.; Hassidim, A.; and Matias, Y. 2022. TRUE: Re-evaluating Factual Consistency Evaluation. 3905–3920.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Köpf, A.; Kilcher, Y.; von Rütte, D.; Anagnostidis, S.; Tam, Z.-R.; Stevens, K.; Barhoum, A.; Duc, N. M.; Stanley, O.; Nagyfi, R.; ES, S.; Suri, S.; Glushkov, D.; Dantuluri, A.; Maguire, A.; Schuhmann, C.; Nguyen, H.; and Mattick, A. 2023. OpenAssistant Conversations – Democratizing Large Language Model Alignment. arXiv:2304.07327.

Liu, N.; Zhang, T.; and Liang, P. 2023. Evaluating Verifiability in Generative Search Engines. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7001–7025. Singapore: Association for Computational Linguistics.

Ni, J.; Qu, C.; Lu, J.; Dai, Z.; Abrego, G. H.; Ma, J.; Zhao, V.; Luan, Y.; Hall, K.; Chang, M.-W.; et al. 2022. Large Dual Encoders Are Generalizable Retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9844–9855.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Gray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe,

R. 2022. Training language models to follow instructions with human feedback. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Pillutla, K.; Swayamdipta, S.; Zellers, R.; Thickstun, J.; Welleck, S.; Choi, Y.; and Harchaoui, Z. 2021. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Rubin, S. J. A. O.; Yoran, O.; Wolfson, T.; Herzig, J.; and Berant, J. 2022. QAMPARI:: An Open-domain Question Answering Benchmark for Questions with Many Answers from Multiple Paragraphs. *arXiv preprint arXiv:2205.12665*.

Stelmakh, I.; Luan, Y.; Dhingra, B.; and Chang, M.-W. 2022. ASQA: Factoid Questions Meet Long-Form Answers. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 8273–8288. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, H.; Wang, R.; Mi, F.; Deng, Y.; Wang, Z.; Liang, B.; Xu, R.; and Wong, K.-F. 2023a. Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12047–12064. Singapore: Association for Computational Linguistics.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023b. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E. H.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Wu, D.; Zhang, J.; and Huang, X. 2023. Chain of Thought Prompting Elicits Knowledge Augmentation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 6519–6534. Toronto, Canada: Association for Computational Linguistics.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2023. A Survey of Large Language Models. arXiv:2303.18223.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*.

Appendix

A. Prompt Case Studies

We report a real-world case of the 1-shot CoT prompt in Table 7. The prompt instructs LLMs to generate a fluent and coherent answer for the given question by mimicking the Human Thought Process demonstrated in the 1-shot CoT demonstration.

In addition, we report a real-world prompt used by the LLM-based CIM in Table 8. The prompt is used to instruct LLMs to add proper citations to the given statement by picking up the instruction and the given in-context learning example.

The ASQA dataset is used for the above two case studies.

B. Benchmark Details

ALCE is the first reproducible benchmark for automatically evaluating LLMs' generations with citations. Gao et al. (2023) randomly select 1,000 examples from ASQA (Stelmakh et al. 2022), QAMPARI (Rubin et al. 2022), and ELI5 (Fan et al. 2019) to build the benchmark, which covers different types of questions. The main purpose of ALCE is to automatically evaluate long text generations with citations. Additionally, it allows LLMs to cite multiple documents for individual statements.

Three evaluation metrics, i.e., Fluency, Correctness, and Citation Quality, are defined to comprehensively evaluate the quality of text generation with citations (Gao et al. 2023):

- **Fluency**: measure if LLMs can generate fluent and coherent answers.
- **Correctness**: measure if LLMs can generate correct answers that cover all aspects of interest.
- **Citation Quality**: measure if LLMs can generate answers that are well supported by the cited documents and no irrelevant documents are cited.

Gao et al. (2023) propose to use MAUVE (Pillutla et al. 2021) to measure the Fluency, and a natural language inference model (Honovich et al. 2022) to evaluate the Citation Quality, and design dataset-specific Correctness evaluation method. They also prove that the three evaluation metrics together contribute to a robust evaluation, and their human evaluations show a strong correlation with the three metrics.

C. LLM Details

We report more details of the six open-source LLMs in Table 9. The LLaMA-13B and LLaMA-33B models are not trained to align with human intentions through supervised fine-tuning. In contrast, the other four supervised fine-tuned models show promising abilities in understanding human instructions and generating responses accordingly.

D. Prompt Strategy Details

Gao et al. (2023) formalize the definitions of the four prompt strategies. Based on their definitions, we define our CoT prompting strategies. The main difference between their and our strategies lies in that their prompt contains 2 demonstrations for in-context learning (Brown et al. 2020); in contrast, we prepend only one CoT (a.k.a. 1-shot CoT) demonstration to fit LLMs' context window size. Additionally, we claim that the token length of our 1-shot CoT demonstration is always smaller than the token length of their 2 demonstrations, making our prompts more safely to fit the window size limitation.

The other settings of the prompt strategies are shown below:

- **VANILLA**. Provide LLMs with the top-3 retrieved documents and instruct LLMs to cite them accordingly. For fair comparisons, we use the retrieved documents included in the ALCE benchmark
- **SUMM**. First generate a summary for each of the top-10 retrieved documents, then provide LLMs with these summaries instead of the full texts and instruct LLMs to cite them accordingly. For fair comparisons, we use the summaries included in the ALCE benchmark.
- **SNIPPET**. First extract a snippet from each of the top-10 retrieved documents, then provide LLMs with these snippets instead of the full texts and instruct LLMs to cite them accordingly. For fair comparisons, we use the snippets included in the ALCE benchmark.
- **ORACLE**. This strategy provides LLMs with three gold documents instead of retrieved ones. However, the original datasets do not contain such gold documents. Gao et al. (2023) propose to approximate five gold documents by running a document filtering algorithm on the top-100 retrieved documents. For fair comparisons, we use the approximated gold documents included in the ALCE benchmark.

The main reason for using summaries and snippets is that they significantly reduce the document length. For example, they reduce document length by approximate $6\times$ on ASQA (Gao et al. 2023). However, they are lossy compressions.

Table 7: A real-world case of the 1-shot CoT prompt.

<Below is the prompt text of a 1-shot CoT demonstration.>
Instruction: You are required to write a high-quality answer for the given question using only the provided search documents and cite them properly. When citing several search results, use the format such as[1][2], [1][3], [2][3], or [1][2][3]. Note that you must cite at least one document in each sentence. In other words, you are forbad to write sentences without citing any document. Moreover, you should add the citation symbols inside each sentence, especially before the period (punct.).

Question: Who set the record for longest field goal?

Document [1](Title: Field goal): toward its own end. The longest field goal kick in NFL history is 64 yards,...
Document [2](Title: Field goal range): 35 and 40 yard lines (closer in a crosswind) often will go for the more ...
Document [3](Title: Field goal): both end zones) is only 66 yards. Scaccia, while playing indoor football, ...

Reasoning Steps:
Step1: I judge Document [1] and Document [2] are relevant to the question.
Step2: According to Document [1], I write an answer statement and cite the Document, namely '''The record for the longest field goal in an NFL game was set by Matt Prater at 64 yards [1].'''
Step3: According to Document [2], I write an answer statement and cite the Document, namely '''But the record for the longest field goal at any level was 69 yards, kicked by collegiate kicker Ove Johansson in a 1976 Abilene Christian University football game against East Texas State University [2].'''
Step4: I merge, rank and concatenate the above two answer statements to obtain a fluent and coherent answer.

Answer: The record for the longest field goal in an NFL game was set by Matt Prater at 64 yards [1]. But the record for the longest field goal at any level was 69 yards, kicked by collegiate kicker Ove Johansson in a 1976 Abilene Christian University football game against East Texas State University [2].

<Below is the prompt text to guide LLMs to answer the given question.>
Instruction: You are required to write a high-quality answer for the given question using only the provided search documents and cite them properly. When citing several search results, use the format such as[1][2], [1][3], [2][3], or [1][2][3]. Note that you must cite at least one document in each sentence. In other words, you are forbad to write sentences without citing any document. Moreover, you should add the citation symbols inside each sentence, especially before the period (punct.).

Question: Who has the highest goals in world football?

Document [1](Title: Argentina{Brazil football rivalry): "Football Player of the Century", by IFFHS International ...
Document [2](Title: Godfrey Chitalu): have beaten Gerd Müller's record of 85 goals in a year, the Football Association ...
Document [3](Title: Godfrey Chitalu): highest official tally claimed by a national football association. Chitalu ...

Answer:

Table 8: A real-world prompt used by the LLM-based CIM.

```
<Below is the prompt text of in-context learning.>
Instruction: You are required to cite relevant documents for the given sentence, where
the cited documents can support the factual claim of the sentence. If you cite several
documents, use the format such as [1][2], [1][3], [2][3] or [1][2][3]. Moreover, you
should add the citation symbols inside each sentence, especially before the period
(punct.).

Document [1] (Title: Cherrapunji): Cherrapunji Cherrapunji (; with the native name Sohra
being more commonly used,...
Document [2] (Title: Cherrapunji): Radio relay station known as Akashvani Cherrapunji.
It broadcasts on FM frequencies ...
Document [3] (Title: Mawsynram): Mawsynram Mawsynram () is a village in the East Khasi
Hills district of Meghalaya state in ...

Several places on Earth claim to be the most rainy, such as Lloró, Colombia, which
reported an average annual rainfall of 12,717 mm between 1952 and 1989, and López de
Micay, Colombia, which reported an annual 12,892 mm between 1960 and 2012. ---> Several
places on Earth claim to be the most rainy, such as Lloró, Colombia, which reported an
average annual rainfall of 12,717 mm between 1952 and 1989, and López de Micay,
Colombia, which reported an annual 12,892 mm between 1960 and 2012 [3].


<Below is the prompt text to instruct LLMs to add citations to the given statement.>
Instruction: You are required to cite relevant documents for the given sentence, where
the cited documents can support the factual claim of the sentence. If you cite several
documents, use the format such as [1][2], [1][3], [2][3] or [1][2][3]. Moreover, you
should add the citation symbols inside each sentence, especially before the period
(punct.).

Document [1] (Title: The Sound of Silence): The Sound of Silence "The Sound of Silence",
originally "The Sounds of Silence" ...
Document [2] (Title: Sounds of Silence): Sounds of Silence Sounds of Silence is the
second studio album by Simon & Garfunkel ...
Document [3] (Title: The Sound of Silence): downloadable content for the video game,
"Rock Band 4". The Disturbed version was ...

The original artist of the sound of silence is Simon & Garfunkel. --->
```

Table 9: Details of the six open-source LLMs. "SFT" is short for supervised fine-tuning. "Base" denotes the foundation LLM.

LLM	SFT	Base	URLs
LLaMA-13B	✘	-	https://huggingface.co/huggyllama/llama-13b
LLaMA-2-13B-Chat	✔	LLaMA-2-13B	https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
Vicuna-13B	✔	LLaMA-13B	https://huggingface.co/lmsys/vicuna-13b-v1.5
LLaMA-33B	✘	-	https://huggingface.co/huggyllama/llama-30b
Oasst-33B	✔	LLaMA-33B	https://huggingface.co/OpenAssistant/oasst-sft-6-llama-30b-xor
LLaMA-2-70B-Chat	✔	LLaMA-2-70B	https://huggingface.co/meta-llama/Llama-2-70b-chat-hf