

# AWeCita: Generating Answer with Appropriate and Well-grained Citations Using LLMs

Suifeng Zhao<sup>1</sup>, Tong Zhou<sup>2,3</sup>, Zhuoran Jin<sup>2,3</sup>, Hongbang Yuan<sup>2,3</sup>, Yubo Chen<sup>2,3†</sup>, Kang Liu<sup>2,3</sup>, Sujian Li<sup>4</sup>

<sup>1</sup>International School, Beijing University of Posts and Telecommunications, Beijing 100876, China

<sup>2</sup>The Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation,  
Chinese Academy of Sciences, Beijing 100190, China

<sup>3</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China

<sup>4</sup>School of Computer Science, Peking University, Beijing 100871, China

**Keywords:** Question answering; Large language models; Citation; Citation appropriateness; Citation granularity

Citation: Zhao S.F., Zhou T., Jin Z.R., et al.: AWeCita: Generating Answer with Appropriate and Well-grained Citations Using LLMs. *Data Intelligence* 6(4), 1134-1157 (2024). doi: <https://doi.org/10.3724/2096-7004.di.2024.0035>

Submitted: April 09, 2024; Revised: May 10, 2024; Accepted: May 15, 2024

---

## ABSTRACT

Large language models (LLMs) excel in various Natural Language Processing tasks but struggle with hallucinations, leading to potentially misleading responses. Researchers have extensively explored LLMs' citation practices. However, existing efforts often overlook the crucial aspects of the appropriateness and granularity of citation, which are vital for mitigating hallucination and enhancing interpretability. To bridge this gap and improve the quality of citations, we propose the Generating Answers with Appropriate and Well-grained Citations using LLMs task (AWeCita), with a focus on citing appropriately with a well granularity. Based on the traditional evaluation metrics of answer accuracy and citation correctness, we introduce two new evaluation metrics, citation appropriateness and citation granularity, to assess LLMs' performance on this task more comprehensively and accurately. We conduct a series of exploratory experiments on ASQA and ELI5 datasets. The experimental results show that, AWeCita outperforms traditional tasks in the metric of citation granularity, most of our methods show a certain advantage incitation appropriateness, however, the improvement towards well-grained citation affects the quote-level citation correctness.

---

---

<sup>†</sup> Corresponding author: Yubo Chen (Email: [yubo.chen@nlpr.ia.ac.cn](mailto:yubo.chen@nlpr.ia.ac.cn); ORCID:0000-0002-5485-9916).

## 1. INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of Natural Language Processing tasks, such as question answering and dialogue system [1, 2, 3]. Despite their impressive performance, LLMs are facing a notable challenge of hallucinations, wherein they may generate responses with misleading or inaccurate factual information [4, 5, 6]. In order to enhance the verifiability and interpretability of the generated responses, one solution is to search a document repository for evidence supporting these responses and cite it appropriately. This practice, known as generating answers with citations, has been extensively explored by numerous scholars [7, 8, 9]. Gao et al. [10] proposed the ALCE benchmark that enables LLMs to automatically generate citations and assess their qualities. Subsequently, researchers have explored different dimensions of this task, such as the comprehensiveness of retrieval information [11], the accuracy of answers [12], and the relevance of the cited documents to the answers [13]. However, while effectively employing externally retrieved documents as cited references, previous works of using citations to enhance interpretability have overlooked the **appropriateness** and the **granularity** of citations.

**The Appropriateness of Citation:** Previous evaluation benchmarks of generation text with citations require every sentence in the response to be cited with one or more evidence regardless of whether it is necessary [10]. In real-world scenarios, it is common that the retrieved documents may not necessarily cover all the knowledge required to generate a response. In other words, it is highly likely that LLMs generate responses using its internal knowledge, which is not encompassed in the retrieved documents. In such cases, citations should not be included.

However, previous citation evaluations have defaulted to the idea that every sentence requires citations. This oversight has side effects in assessing citation correctness. For instance, as shown in Figure 1(a), the sentence “Tied with the Celtics for the most NBA championship titles, the Los Angeles Lakers are another legendary team making waves in basketball history” cannot be factually supported by the retrieved documents, so citations should not be included. However, the traditional evaluation of ALCE [10] assigns the recall of this sentence to 0, incorrectly attributing it to a document selection problem. Therefore, it will be beneficial to prioritize appropriate citation practices.

**The Granularity of Citation:** The cited references not only serve as the foundation for answers but also play a pivotal role in providing interpretability and guiding the reader to trace back to the source of evidence. Nevertheless, previous works often present references at a document-level granularity [10], which is over-coarse-grained, and hinders readers’ efficiency in verification. Conversely, over-fine-grained citations may lead to the insufficiency of supporting citations, making it challenging for readers to trace back effectively. As depicted at the top of Figure 1(b), over-coarse-grained citations contain a considerable amount of redundancy besides the precise evidence “Christine Sinclair is officially the world’s all-time leader for international goals scored for women with 190 goals”. This brings burden to the readers. Conversely, the example in the middle of 1(b) demonstrates an over-fine-grained citation example. Using only the name “Christine Sinclair” as a citation is insufficient for readers to trace back to the source of

**Question:** Which teams have won the most NBA championships?

**AWeCita:** Winning the most NBA championships with 17 NBA championship titles to their name, the Boston Celtics are one of the most successful franchises in NBA history [1]. Tied with the Celtics for the most NBA championship titles, the Los Angeles Lakers are another legendary team making waves in basketball history.

*Note: The second sentence is generated without referencing external documents, so citations should not be included.*

**Traditional Citation:** Winning the most NBA championships with 17 NBA championship titles to their name, the Boston Celtics are one of the most successful franchises in NBA history [1]. Golden State Warriors won Boston Celtics in 2022 [2]. So, Los Angeles Lakers and Golden State Warriors are the most championship winning teams in NBA history [1][2].

**Retrieved Documents:** 📄

**Relevant Document [1]:** Winning the most NBA championships with 17 NBA championship titles to their name, the Boston Celtics are one of the most successful franchises in NBA history. They won their first championship in 1957 and added 16 more in the following years—including 1960, 1961, and 1962—marking the beginning of their dominance. Their most recent NBA title came in 2008 when they defeated the Los Angeles Lakers in the Finals with an incredible season and in-game comeback.

**Interference Document [2]:** Golden State Warriors: the team has won a total of seven league championships. The 2015-16 Golden State Warriors achieved 73 wins and 9 losses, becoming the best single-season regular-season record in NBA history, breaking the 1995-96 Chicago Bulls' record of 72 wins and 10 losses. The team won the NBA championship for the first time by sweeping the Cleveland Cavaliers, 4-0, in the 2018 Finals, and then reached a second consecutive title after going three seasons to win another NBA title in the 2022 Finals with a 4-2 win over the Boston Celtics.

(a)

**Question:** Who has the highest goals in women's world international football?

**Answers&Citations:** 📄

**Over-Coarse-Grained Citation**

**Answer:** Christine Sinclair has the highest goals in women's world international football {1}.

**Citation {1}:** Christine Margaret Sinclair (born June 12, 1983) is a Canadian professional soccer player who plays as a forward for the Portland Thorns of the National Women's Soccer League (NWSL) and, from 2000 until her retirement from international football in 2023, was a member of the Canadian national team. Christine Sinclair is officially the world's all-time leader for international goals scored for women with 190 goals, and is one of the most-capped international soccer players with 331 appearances.

**Over-Fine-Grained Citation**

**Answer:** Christine Sinclair has the highest goals in women's world international football {1}.

**Citation {1}:** Christine Sinclair

**Well-Grained Citation**

**Answer:** Christine Sinclair has the highest goals in women's world international football {1}.

**Citation {1}:** Christine Sinclair is officially the world's all-time leader for international goals scored for women with 190 goals.

(b)

**Figure 1.** Example on citation appropriateness and citation granularity. (a) An example of comparing traditional citation with the new task. (b) Over-coarse-grained, over-fine-grained, and well-grained citation comparison.

knowledge and verify the answer because it is too short. Therefore, well-grained citation contributes to enhancing interpretability and verifiability.

In order to tackle the two aforementioned problems, this paper introduces AWeCita, a task focused on Generating Answers with Appropriate and Well-grained Citations using LLMs. Along-side this task, two evaluation metrics are proposed: **citation appropriateness** and **citation granularity**, besides traditional metrics: answer accuracy and citation correctness. Citation appropriateness evaluates whether the model accurately determines whether to include a citation or not for a sentence. The other proposed metric, citation granularity assesses whether citations are included with the suitable granularity. Specifically, we define a citation as well-grained when the cited quote precisely supports the response without redundancies. We have designed and implemented a series of methods for the AWeCita task, which can be categorized into one-step-citation, including *link-quotation* and *inline-quotation*, as well as a two-step-citation approach *post-quotation*. We conduct experiments on the ASQA [14] and ELI5 [15] datasets, uncovering some insightful phenomena within them.

Our contributions can be summarized as follows. (1) We introduce the AWeCita task, which focuses on citation appropriateness and citation granularity, and propose two evaluation metrics tailored to the task: citation appropriateness, which evaluates the model's decision on whether or not to cite, and citation granularity, which assesses the precision of cited spans. (2) We test the performance of traditional citation baseline ALCE [10], and analyze its shortcoming. (3) We propose three methods tailored to the AWeCita task and conduct experiments accordingly; (4) We draw a series of valuable experimental conclusions, including: AWeCita outperforms traditional tasks in the metric of citation granularity, the improvement towards well-grained citation affects the quote-level citation correctness, and most of our methods show a certain advantage in citation appropriateness.

## **2. RELATED WORK**

**Generate Text with Citation:** In the domain of Question Answering (QA), citations serve as valuable means to enhance reliability and interpretability by providing valuable support and contextual information [16, 17]. The ALCE benchmark [10] proposes the task paradigm and evaluation system for LLMs' citation generation. This benchmark encompasses several citation generation paradigms including vanilla, summaries or snippets, and inline-search. Moreover, it evaluates not only the accuracy of answers through fluency and accuracy, but also the quality of citations. This task presents a notable challenge as there is considerable scope for citation quality enhancement while generating answers. Subsequent studies have explored various approaches to improve citation qualities. Among which, many works focusing on architectural modifications to improve the effectiveness of citations. Sun et al. [18] proposes a new framework VTG, an evolving memory system and a two-tier verification mechanism, thereby improving the reliability and factual grounding of generated content. Besides, fine-tuning is proved to be effective in unleashing models' capabilities. Ye et al. [19] utilizes the approach of self-referencing and iterative refinement to finetune large language models, thereby enhancing their ability to generate factually grounded responses with improved citation accuracy. Although related work has led to advances in the citation quality, the effects on citation appropriateness and citation granularity are remained unclear.

**Perspective of Citation Evaluation:** The ALCE benchmark assesses the adequacy and necessity of citations. Adequacy requires that cited documents must support the answer, while necessity dictates not to cite redundant documents. Yue et al. [20] classifies citations into three types: attributable, extrapolatory, and contradictory, which provides a more granular framework for automatically evaluating the attribution of LLMs. Li et al. [21] focuses on the inevitability of an inadequate knowledge source and enhances the generation reliability by making the model aware of the limitations of its knowledge based on the knowledge graph. Though there hasn't been any prior work proposing an evaluation of knowledge integration and citation granularity, discussions on both topics are starting to emerge. Huang and Chang [22] point out that citation requires both external non-parametric knowledge and internal parametric knowledge, and over-citing may potentially impede models' creativity. Wang et al. [23] has begun to focus on allowing the model to recall fine-grained fragments from coarse-grained documents,

improving the accuracy and efficiency of reference information retrieval and demonstrating the significant benefits of fine-grained reference information in knowledge-intensive tasks. Our work is the first to consider both the appropriateness and a more suitable granularity in LLMs' citation generation tasks, providing valuable insights into the methods and paradigms at the same time.

### 3. PROBLEM FORMULATION

#### 3.1 Task Definition

Given an input question  $q$  and a set of documents  $D = \{d_1, d_2, \dots, d_k\}$ , which are selected from the retrieved documents (it is possible that some of these documents maybe unhelpful). The desired output consists of the following components: (1) the answer sentences  $\{s_1, s_2, \dots\}$ ; (2) the references  $\{r_1^i, r_2^i, \dots, r_m^i, \dots\}$  corresponding to each answer sentence  $s_i$ , with each reference consisting both of the id of the referred document  $d_j \in D$  and the exact quoted span  $q_j \in Q$  within  $d_j$ , i.e.  $r_m^i = (d_j, q_j)$ , where  $Q$  is a set of quoted spans.

It is worthy noting that, the setting for AWeCita task is distinguished from previous work [10, 18] in three aspects. Firstly, previous works requires LLMs to provide citation for any sentence they generate. In contrast, AWeCita requests LLMs to cite if and only if the document set can support the sentence. This means that the references  $r$  should be included exclusively when  $D$  supports the given sentence. Secondly, instead of document-level references, AWeCita offers citations at the better granularity on condition that they can fully support the sentence. In other words, AWeCita annotates the exact quoted span rather than simply marking document ids. Thirdly, traditional citation works always utilize the top  $k$  relevant documents retrieved based on the question  $q$ . Instead, AWeCita intentionally mixes in one or several noisy documents with the aim of assessing the model's ability to accurately identify evidences. To be specific, we mix one or more most irrelevant retrieved documents into the document set  $D$ .

#### 3.2 Datasets

The experiments of the AWeCita task are conducted on two datasets: ASQA and ELI5.

**ASQA:** The ASQA [14] (Answer Summaries for Questions which are Ambiguous) dataset is a long-form question-answering dataset that focuses on ambiguous factoid questions which have different correct answers depending on the interpretation. The questions mainly focus on entities such as people, quantities, and locations. Short answers covering different aspects and gold long answers are provided in this dataset, enabling users to measure the correctness of model generations.

**ELI5:** The ELI5 [15] (Explain Like I'm Five) dataset is also a long-form QA dataset. Unlike ASQA, which contains clear entities in the answer, ELI5 mainly focuses on more comprehensive questions like asking how/why/what, which requires multiple knowledge sources to answer. To aid in answering these questions,

the dataset offers a substantial collection of web documents and sample answers, allowing users to better measure the correctness of models' responses.

### 3.3 Evaluation

Four evaluation metrics are employed to assess the quality of the answers and citations. Among them, *answer accuracy* and *citation correctness* follow on with the previous work [10].

#### 3.3.1 Citation Appropriateness (*app*)

The paper specifically designs a metric *app* ( $D, s, c$ ) to evaluate the ability of the model on determining whether to cite or not for each generated sentence, named Citation Appropriateness.

Firstly, given a sentence  $s$ , this metric assesses whether the document set  $D = \{d_1, d_2, \dots\}$  as a whole can sufficiently support it. We concatenate all the documents in  $D$  to form one combined-document, i.e.  $ent(\text{concat}(d_i \in D), s)$  in Eq.(1), and measure whether it logically entails  $s$ , using a Natural Language Inference (NLI)[24] model.

$$ent(\text{concat}(d_i \in D), s) = \begin{cases} 1, & \text{if } \text{concat}(d_i \in D) \text{ entails } s \\ 0, & \text{if } \text{concat}(d_i \in D) \text{ cannot entail } s \end{cases} \quad (1)$$

Secondly, the model's ability to cite appropriately is evaluated by comparing its citation decision  $c$  with the entailment result  $ent(\text{concat}(d_i \in D), s)$  obtained above. The expectation is that the model should include citations only for sentences supported by the external documents, at the same time avoid including citations for those that are irrelevant to the external documents. The validation process is outlined in Eq (2).

$$app(D, s, c) = val(ent(\text{concat}(d_i \in D), s), c) \quad (2)$$

where *app* ( $D, s, c$ ) represents the citation appropriateness for each sentence  $s$ ,  $D$  is the document set,  $c$  is a boolean value, representing whether citations are included in  $s$ ; *ent* returns a boolean value, indicating whether the combined passage logically supports the sentence,  $d_i$  denotes a document in  $D$ ; *val* returns a boolean value that represents whether the entailment result  $ent(\text{concat}(d_i \in D), s)$  matches the model's citation decision  $c$ .

Finally, we average through the entire dataset to calculate the citation appropriateness for each sentence to obtain the citation appropriateness for a model on a specific dataset.

#### 3.3.2 Citation Correctness (*corr (doc)* and *corr (quote)*)

The evaluation of citation correctness includes two metrics: document-level correctness and quote-level correctness. Document-level correctness evaluates whether the cited document set  $D$  is cited correctly by

measuring whether it supports the sentence  $s$  (Eq.(3)), while quote-level correctness evaluates whether the quoted span set  $Q$  is the correct span set to sufficiently support  $s$  (Eq.(4)).

$$\text{corr}(s, D) = \text{ent}(\text{concat}(d_i), s) \quad (3)$$

$$\text{corr}(s, Q) = \text{ent}(\text{concat}(q_i), s) \quad (4)$$

where  $d_i$  is a document in the cited document set  $D$ ,  $q_i$  is a quoted span in  $Q$ ,  $\text{ent}(\text{concat}(d_i), s)$  is a boolean value, indicating whether the concatenation of  $d_i$  logically entails the sentence  $s$ , and similarly  $\text{ent}(\text{concat}(q_i), s)$  indicates whether the concatenation of  $q_i$  entails  $s$ .

Finally, we can obtain the document-level citation correctness and quote-level citation correctness for a model on a specific dataset. This is achieved by respectively averaging  $\text{corr}(s, D)$  and  $\text{corr}(s, Q)$  across the entire dataset.

This assessment is conducted exclusively only when citations are needed and the model has included citations in the sentence.

### 3.3.3 Citation Granularity (*gran*)

This metric aims to determine whether the model quotes the most concise spans to support the generated sentence, and score the cited quotes accordingly.

Given a sentence  $s$  and the generated quoted span set  $Q = \{q_1, q_2, \dots\}$ , we begin by breaking down all the quotes  $q \in Q$  into clauses according to punctuation such as commas and periods. The total number of clauses is noted as  $|Q|$ . Then, we need to find a gold clause combination set  $Q_s$  from  $Q$ , which can sufficiently support  $s$  with the minimum number of clauses. The number of clauses in the  $Q_s$  is noted as  $|Q_s|$ . The citation granularity is defined using  $\text{gran} = \frac{|Q_s|}{|Q|}$ .

To find the gold clause combination  $Q_s$ , and further calculate the citation granularity, we employ an iterative search process based on a pruning strategy, which is shown in Algorithm1. Specifically, we first define  $Q_s$  to store the gold combination, and  $|Q_s|$  to monitor the number of clauses in it. Second, for a given set  $Q$ , we begin by iterating the number of clauses from  $|Q|$  downwards. At each iteration, we enumerate all possible clause combinations corresponding to the current number, ensuring that clauses selected from the same quote are continuous, using the function *generate\_combinations* ( $Q, \text{num}$ ). Then, we assess whether these combinations can logically entail  $s$  using  $\text{ent}(\text{comb}, s)$ . We assume that if a larger set cannot entail sentence  $s$ , then none of its subsets can entail it. Conversely, if a small subset entails the sentence  $s$ , and even with additional information attached to it, it still entails it. Based on this assumption, if a combination ails to entails, we prune all subsets of that combination, using *prune\_subsets* ( $\text{comb}$ ). In contrast, if a combination successfully entails the  $s$ , we continue to explore smaller span combinations until no smaller combination can be found. Finally, we can obtain the citation granularity:  $\text{gran} = \frac{|Q_s|}{|Q|}$  and the gold clause combination  $Q_s$ .

---

**Algorithm 1.** Granularity Calculation using Pruning Strategy.

---

$\{Q_s\}$ : the gold clause combination set;  $|Q_s|$ : the number of clauses in  $Q_s$ ;  $|Q|$ : the number of clauses in the given quotation set  $Q$ ;  $gran$ : the granularity of  $Q$

**Input:**  $Q$  {Set of quotes}

$Q_s \leftarrow \text{null}$

$|Q_s| \leftarrow \infty$

$|Q| \leftarrow \text{count clauses } (Q)$

**for** num **in** range ( $|Q|$ , 0,  $\square 1$ ) **do**

combs  $\leftarrow$  generate combinations ( $Q$ , num)

**for** comb **in** combs **do**

**if** ent (comb,  $s$ ) **then**

gold  $\leftarrow$  min (gold, count clauses (comb))

$Q_s \leftarrow \text{comb}$

**else**

prune subsets (comb)

**end if**

**end for**

**end for**

$gran \leftarrow \frac{|Q_s|}{|Q|}$

**Output:**  $gran, Q_s$

---

This assessment occurs exclusively when the citation is deemed correct during the calculation of quote-level citation correctness (i.e.  $corr(s, Q) = 1$ ).

The best citation quality  $bestc$  is achieved when a citation exhibits both high quote-level citation correctness and high granularity simultaneously, indicated by a high value of  $corr(s, Q) * gran$ , which means  $bestc = \max(corr(s, Q) * gran)$ .

### 3.3.4 Answer Accuracy ( $acc$ )

This metric evaluates whether the model's response to the question is accurate. In detail, a comparison is conducted between the model-generated answers and manually annotated answers. The evaluation methods vary according to the datasets, including string matching and a NLI model.

4. METHODOLOGY

We employ various methodologies within the AWeCita to assess the performance of different citation strategies. Initially, the model is presented with a question and multiple retrieved documents of that question. Afterwards, the model is prompted to generate an answer with appropriate citations and in proper granularity. With the aim of examining the quality of answers and citations in various scenarios, we employ different baseline methods. These include *Link Quotation* and *Inline Quotation*, both of which are one-step citation methods, as well as *Post Quotation*, which involves a two-step citation generation process.

The **link quotation** is in the form of an appendix, which is the most direct and close to the human reading habit. Then to better fit the reading habits of the model, we further designed **inline quotation**, which uses the latex form of quotation. In addition, we have designed the **post quotation** method, which is edited on the basis of the traditional citation method, in order to consider the balance between citation correctness and citation granularity.

**Link Quotation:** As illustrated at the top of Figure 2, link quotation involves creating reference links alongside the answer, with the relevant quoted spans listed subsequently. Specifically, given a question,

**Question:** Who has the highest goals in women’s world international football?

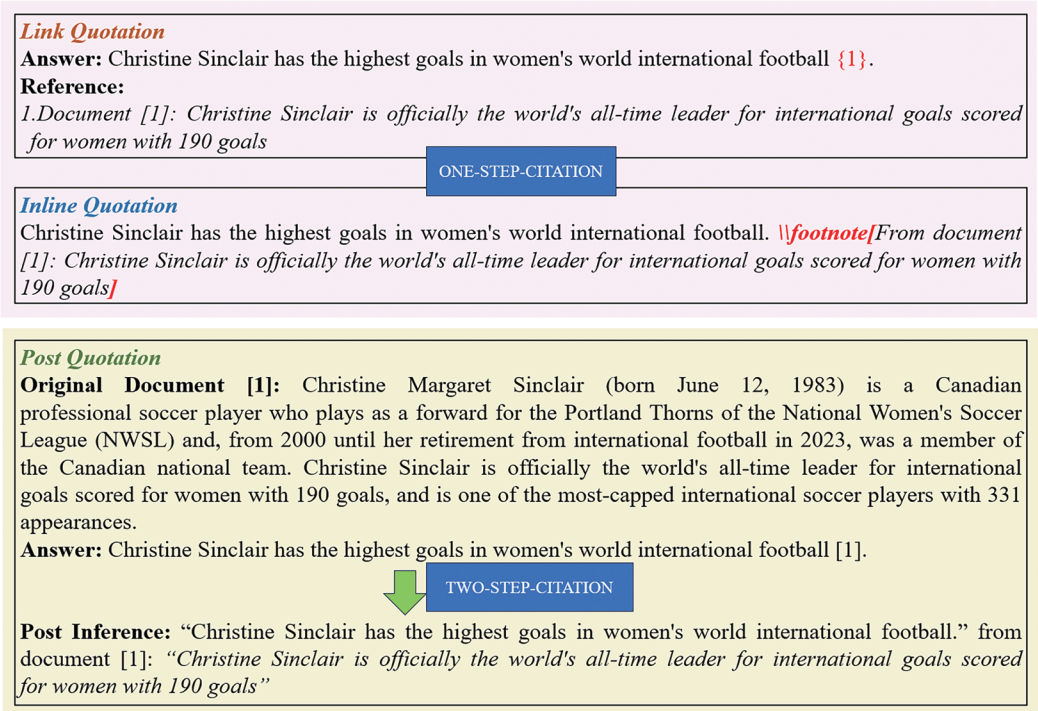


Figure 2. Comparison between different citation methods: Link Quotation, Inline Quotation, and Post Quotation.

this method generates an answer and includes a link for each sentence which requires citation. Simultaneously, the model provides a list of quoted spans corresponding to each link. These quoted spans highlight specific segments from the source documents that contribute to the generated answer. By including both cited links and quoted spans, the Link Quotation method enhances the transparency and credibility of the generated content. It enables readers to directly access the exact quoted span from the cited link, facilitating verification of the information.

**Inline Quotation:** While link quotation presents citations in a clear format, it introduces a time delay by directing users to a list of quoted spans and involves various types of citation marks. These factors may have a negative impact on the model's generation process. To address potential challenges, we propose inline quotation, which provides a more immediate citation and simplifies the format of citations. As depicted in the middle of Figure 2, inline quotation visibly presents the quoted reference spans alongside the corresponding answers. This allows the source of information to be referenced immediately after the answer, providing readers with direct access to the cited sources.

**Post Quotation:** In contrast to link quotation and inline quotation, which generate responses and quotations simultaneously, the post-quotation method involves a two-step process for generating answers, as depicted in the bottom of Figure 2. Firstly, the model is provided with a question and the retrieved documents, and it is prompted to generate an answer while citing the supporting documents, akin to the traditional citation method. Subsequently, the model analyzes the provided answer alongside the reference documents to generate more precise quotations. This process ensures proper attribution and support for the generated content.

## **5. EXPERIMENTS AND RESULTS**

Series of experiments are conducted to explore how to achieve optimal results for the AWeCita task on the four metrics: answer accuracy, citation appropriateness, citation correctness (including document correctness and quote correctness), and citation granularity.

The experimental setup involves the utilization of two commonly used QA datasets, namely ASQA [14] and ELI5 [15]. We employed a diverse set of models, including LLaMA2-7b-Chat, LLaMA-13b-Chat, LLaMA2-70b-Chat [25], and a closed-source model gpt-3.5-turbo-0125. These models were selected to encompass a broad range of scales, incorporating both open-source and closed-source variants, which helps to investigate the impact of model scale on metrics and determine whether closed-source models exhibit superior performance. To make fair comparison, this paper follows the previous work and uses different metrics to evaluate answer accuracy in two datasets, respectively. Specifically, we use string matching to compare answers in the ASQA dataset, and use a Natural Language Inference (NLI) [24] model for ELI5 dataset. We also use a NLI model to check if a supporting reference can necessarily entail a sentence. For decoding hyperparameter settings, a temperature of 0.5 was used, with maximum input and output token lengths set to 4096. We maximize the values of shot and seed to accommodate the token length limit of 4096 characters. Through this comprehensive parameter configuration, we conducted

experiments to evaluate the performance and effectiveness of the models across different datasets and parameter combinations.

Our experiments cover four baselines. The first is the traditional baseline **ALCE** [10], which generates document numbers as citations alongside the answer in one step [10]. The other three are the proposed baselines: **Link Quotation**, **Inline Quotation**, and **Post Quotation**. Given a question with the top 5 relevant documents, each of the methods requires the models to generate answers and citations based the inputs. **Link Quotation** involves appending a link alongside the answer, labeled with a numerical index in the form of *{index}*, followed by a list of relevant quoted spans corresponding to each link quotation mark. The output is structured as *{index}: Document [document number]: quoted span*. **Inline Quotation** requires the model to quote the spans right after the answers, following the format: *answer \\footnote {From document {document number} quoted span}*. **Post Quotation** requires a two-step generation. First, generate answer given the question and documents. Then correspond the answer with the exact spans.

### 5.1 Results and Analysis

Tables 1 and 2 display the evaluation results of different models across four baselines (ALCE, Link Quotation, Inline Quotation, and Post Quotation) on the two datasets (ASQA and ELI5). Our key findings are as follows.

#### 5.1.1 Overall Performance

**Our proposed methods significantly outperform ALCE in granularity.** In various experimental settings, the granularity scores of ALCE typically range from 30 to 40 points, whereas our methods can achieve granularity scores as high as 70 to 80 points. Under the same experimental conditions, our methods exhibits a notable improvement of 10 to 40 points compared to ALCE, showcasing significant advantages and enhancements. This demonstrates that reasonable granularity generation can be enhanced by a appropriate prompt instructions.

**The improvement towards well-grained citation affects the quote-level citation correctness.** When comparing the metric of document-level citation correctness, there are instances where our methods achieve better scores. However, there are also instances where ALCE achieves better scores. This suggests that document-level citation correctness is influenced more by the capabilities of the model rather than the specific method employed. Therefore, we can conclude that while the models' ability to trace back to the original reference document performs well, accurately locating the exact span within the document without losing essential supporting details remains challenging.

**Most of our methods show a certain advantage in citation appropriateness.** In addition to the influence of instructions, citation appropriateness is more influenced by the ability of the model, showing that model's use of internal knowledge is still dependent on its own capabilities. Therefore, LLMs need to be capable enough to reason based on retrieved documents and their own knowledge, incorporating the problem, and thus determining the opportune moment to make citations.

**Table 1.** Main result on dataset ASQA, covering four baselines: ALCE, representing the conventional citation approach for comparison; alongside our proposed methodologies—Link Quotation, Inline Quotation, and Post Quotation.

|                                     | Answer |       | Citation |            |              |       |
|-------------------------------------|--------|-------|----------|------------|--------------|-------|
|                                     | length | acc   | app      | corr (doc) | corr (quote) | gran  |
| LLaMA2-7B-Chat ALCE                 | 50.73  | 28.39 | 61.12    | 56.47      | 56.47        | 35.27 |
| LLaMA2-7B-Chat Link Quotation       | 43.67  | 28.08 | 60.54    | 51.49      | 30.79        | 53.21 |
| LLaMA2-7B-Chat Inline Quotation     | 64.98  | 31.68 | 64.66    | 57.94      | 32.07        | 77.64 |
| LLaMA2-7B-Chat Post Quotation       | 35.91  | 24.06 | 59.37    | 61.02      | 35.63        | 60.75 |
| LLaMA2-13B-Chat ALCE                | 51.32  | 32.90 | 65.60    | 50.88      | 50.88        | 36.38 |
| LLaMA2-13B-Chat Link Quotation      | 42.46  | 30.85 | 64.81    | 54.33      | 35.81        | 44.85 |
| LLaMA2-13B-Chat Inline Quotation    | 58.28  | 34.20 | 69.52    | 61.38      | 34.25        | 82.78 |
| LLaMA2-13B-Chat Post Quotation      | 42.40  | 29.49 | 67.80    | 57.89      | 25.32        | 68.62 |
| LLaMA2-70B-Chat ALCE                | 74.50  | 38.30 | 74.46    | 71.49      | 71.49        | 36.98 |
| LLaMA2-70B-Chat Link Quotation      | 68.90  | 38.05 | 60.65    | 68.47      | 21.28        | 70.99 |
| LLaMA2-70B-Chat Inline Quotation    | 72.94  | 37.47 | 75.20    | 73.64      | 46.74        | 79.27 |
| LLaMA2-70B-Chat Post Quotation      | 54.18  | 35.38 | 75.04    | 76.68      | 31.46        | 65.57 |
| gpt-3.5-turbo-0125 ALCE             | 48.15  | 34.24 | 81.92    | 86.04      | 86.04        | 31.70 |
| gpt-3.5-turbo-0125 Link Quotation   | 49.88  | 34.80 | 45.30    | 78.47      | 44.59        | 69.66 |
| gpt-3.5-turbo-0125 Inline Quotation | 66.73  | 35.44 | 76.98    | 85.70      | 42.73        | 82.62 |
| gpt-3.5-turbo-0125 Post Quotation   | 44.21  | 33.67 | 82.71    | 90.19      | 50.83        | 67.04 |

Note: Here, *length* us to note the answer length, *acc* to measure the answer accuracy. And *app*, *corr (doc)*, *corr (quote)*, *gran* evaluates citation appropriateness, document-level citation correctness, quote-level citation correctness, and citation granularity respectively. In this table, various colors are employed to distinguish between different methods, with lighter to darker colors in the same color scheme indicating increasing model capability.

**Table 2.** Main result on dataset ELI5, covering four baselines: ALCE, representing the conventional citation approach for comparison; alongside our proposed methodologies—Link Quotation, Inline Quotation, and Post Quotation.

|                                     | Answer |       | Citation |            |              |       |
|-------------------------------------|--------|-------|----------|------------|--------------|-------|
|                                     | length | acc   | app      | corr (doc) | corr (quote) | gran  |
| LLaMA2-7B-Chat ALCE                 | 80.78  | 11.0  | 45.92    | 46.53      | 46.53        | 36.89 |
| LLaMA2-7B-Chat Link Quotation       | 116.67 | 13.11 | 45.85    | 33.94      | 11.30        | 47.08 |
| LLaMA2-7B-Chat Inline Quotation     | 102.06 | 11.67 | 48.64    | 38.70      | 17.81        | 76.96 |
| LLaMA2-7B-Chat Post Quotation       | 49.51  | 9.83  | 46.52    | 49.45      | 31.69        | 41.41 |
| LLaMA2-13B-Chat ALCE                | 82.54  | 12.61 | 44.63    | 42.84      | 42.84        | 36.83 |
| LLaMA2-13B-Chat Link Quotation      | 112.26 | 14.05 | 46.48    | 38.17      | 20.56        | 57.32 |
| LLaMA2-13B-Chat Inline Quotation    | 90.87  | 13.00 | 46.04    | 51.83      | 24.90        | 77.27 |
| LLaMA2-13B-Chat Post Quotation      | 62.02  | 11.42 | 45.21    | 50.46      | 21.81        | 48.18 |
| LLaMA2-70B-Chat ALCE                | 101.00 | 12.48 | 56.61    | 60.68      | 60.68        | 38.00 |
| LLaMA2-70B-Chat Link Quotation      | 141.41 | 15.08 | 56.07    | 62.29      | 25.37        | 61.02 |
| LLaMA2-70B-Chat Inline Quotation    | 108.07 | 13.40 | 54.69    | 59.20      | 32.48        | 76.14 |
| LLaMA2-70B-Chat Post Quotation      | 59.25  | 10.32 | 58.41    | 67.58      | 34.75        | 48.81 |
| gpt-3.5-turbo-0125 ALCE             | 76.27  | 11.43 | 60.56    | 67.80      | 67.80        | 36.12 |
| gpt-3.5-turbo-0125 Link Quotation   | 97.25  | 15.03 | 56.30    | 73.88      | 39.31        | 60.36 |
| gpt-3.5-turbo-0125 Inline Quotation | 91.97  | 12.07 | 65.20    | 72.48      | 35.98        | 78.93 |
| gpt-3.5-turbo-0125 Post Quotation   | 60.75  | 10.87 | 64.25    | 78.90      | 44.01        | 55.10 |

**Capable models consistently demonstrate superior performance**, across all methods and metrics, as shown in table 1 and table 2. More specifically, we observe an overall enhancement in the scores of answer accuracy, citation appropriateness, citation correctness, and citation granularity from LLaMA2-7B-Chat to LLaMA2-13B-Chat, and further to LLaMA2-70B-Chat and gpt-3.5-turbo-0125. This phenomenon shows that larger models demonstrate higher answer accuracy and better citation quality, underscoring the fact that citation tasks in the QA domain are indicative of LLMs' capabilities.

### 5.1.2 Comparison of Different Quotation Methods

**Link Quotation:** While offering promising citation granularity by improving approximately 10-30 points compared to ALCE, link quotation has its weaknesses. Firstly, this citation format tends to generate more continuous citation marks at the end of sentences, thereby reducing granularity performance. Second, this format is hard for model to adapt, leading to a potential confusion between citation numbers and document indices, causing a limited performance of citation correctness on both document-level and quote-level, especially on bigger models like LLaMA2-70B-Chat and gpt-3.5-turbo-0125. Therefore, refining the format to address this limitation is important for improving citation quality.

**Post Quotation:** The core idea of post-quotation is to modify based on ALCE's generated answers, improving granularity scores by 20-30 points. But since the model is more adapted to generating responses and citations in one step form rather than two steps, the post quotation method does not perform as well as inline quotation in terms of granularity. When it comes to the quote-level citation correctness, post quotation does not present an obvious advantage, showing a dominant rely of it on models' capabilities.

**Inline Quotation:** Inline quotation exhibits superior performance in terms of citation granularity, indicating the model's innate adaptability to this format. However, it falls short in consistently showcasing the advantages of the citation correctness metric, which necessitates a broader display of the model's capabilities. This points to the fact that mining the adaptability of the model to the format is a way to increase citation granularity.

## 5.2 Ablation Study

In this section, we perform ablation study to test the influence of two factors: the number of interfering documents, and the granularity of the reference documents. The results are displayed in tables 3 and 4, and we derived several observations from it.

### 5.2.1 Appropriateness Analysis

The hyperparameter configurations of the appropriateness analysis includes setting the total document number provided for each question to 5, with replacement document number ranging in 1, 3, 5. The hyperparameter 1, 3, 5 replacements are chosen to simulate real search scenarios where a small number of documents are irrelevant, more than half of the documents are irrelevant, and no documents are relevant at all. More concretely, when replacement document number is 1, it means to replace the most

**Table 3.** Ablation experiment - number of interfering documents (on ASQA), covering two baselines: ALCE, representing the conventional citation approach for comparison; alongside our proposed method—Inline Quotation.

|                                               | Answer |       | Citation |            |              |       |
|-----------------------------------------------|--------|-------|----------|------------|--------------|-------|
|                                               | length | acc   | app      | corr (doc) | corr (quote) | gran  |
| LLaMA2-7B-Chat ALCE (1-noise)                 | 50.73  | 28.39 | 61.12    | 56.47      | 56.47        | 35.27 |
| LLaMA2-7B-Chat Inline Quotation (1-noise)     | 64.98  | 31.68 | 64.66    | 57.94      | 32.07        | 77.64 |
| gpt-3.5-turbo-0125 ALCE (1-noise)             | 48.15  | 34.24 | 81.92    | 86.04      | 86.04        | 31.70 |
| gpt-3.5-turbo-0125 Inline Quotation (1-noise) | 66.73  | 35.44 | 76.98    | 85.70      | 42.73        | 82.62 |
| LLaMA2-7B-Chat ALCE (3-noise)                 | 51.78  | 20.63 | 48.67    | 55.95      | 55.95        | 40.49 |
| LLaMA2-7B-Chat Inline Quotation (3-noise)     | 64.39  | 23.87 | 51.19    | 61.34      | 32.21        | 73.88 |
| gpt-3.5-turbo-0125 ALCE (3-noise)             | 42.75  | 26.29 | 69.65    | 78.94      | 78.94        | 39.62 |
| gpt-3.5-turbo-0125 Inline Quotation (3-noise) | 59.62  | 27.13 | 66.97    | 82.93      | 41.46        | 83.00 |
| LLaMA2-7B-Chat ALCE (5-noise)                 | 50.02  | 14.22 | 33.61    | 62.74      | 62.74        | 42.90 |
| LLaMA2-7B-Chat Inline Quotation (5-noise)     | 62.45  | 15.90 | 36.75    | 63.57      | 27.45        | 67.03 |
| gpt-3.5-turbo-0125 ALCE (5-noise)             | 38.26  | 17.28 | 48.02    | 77.70      | 77.70        | 37.79 |
| gpt-3.5-turbo-0125 Inline Quotation (5-noise) | 59.38  | 20.57 | 50.82    | 77.11      | 31.25        | 84.89 |

Note: Here, k-noise denotes the interference documents included in the document set for a question.

relevant document (the first in the top 100 retrieved documents) into the last relevant one (the last in the top 100 retrieved documents). Similarly, when replace document number is 3, we replace the first, second, and third documents in the top 100 retrieved documents into the last three documents. And we replace the first five documents into the last five documents when replace document number is 5.

**Table 4.** Ablation experiment - granularity of reference documents (on ASQA), covering two baselines: ALCE, representing the conventional citation approach for comparison; alongside our proposed method—Inline Quotation.

|                                                         | Answer |       | Citation |            |              |       |                   |
|---------------------------------------------------------|--------|-------|----------|------------|--------------|-------|-------------------|
|                                                         | length | acc   | app      | corr (doc) | corr (quote) | gran  | corr (quote)*gran |
| LLaMA2-7B-Chat ALCE (coarse-paragraph)                  | 24.51  | 23.38 | 41.72    | 85.28      | 85.28        | 20.52 | 17.50             |
| LLaMA2-7B-Chat Inline Quotation (coarse-paragraph)      | 61.78  | 30.68 | 43.05    | 65.16      | 36.31        | 75.46 | 27.40             |
| gpt-3.5-turbo-0125 ALCE (coarse-paragraph)              | 35.29  | 32.94 | 57.42    | 90.61      | 90.61        | 15.48 | 14.03             |
| gpt-3.5-turbo-0125 Inline Quotation (coarse-paragraph)  | 63.25  | 34.50 | 49.93    | 82.03      | 44.57        | 81.05 | 36.12             |
| LLaMA2-7B-Chat ALCE (default-paragraph)                 | 50.73  | 28.39 | 61.12    | 56.47      | 56.47        | 35.27 | 19.92             |
| LLaMA2-7B-Chat Inline Quotation (default-paragraph)     | 64.98  | 31.68 | 64.66    | 57.94      | 32.07        | 77.64 | 24.90             |
| gpt-3.5-turbo-0125 ALCE (default-paragraph)             | 48.15  | 34.24 | 81.92    | 86.04      | 86.04        | 31.70 | 27.27             |
| gpt-3.5-turbo-0125 Inline Quotation (default-paragraph) | 66.73  | 35.44 | 76.98    | 85.70      | 42.73        | 82.62 | 35.30             |
| LLaMA2-7B-Chat ALCE (sentence)                          | 66.89  | 30.27 | 52.92    | 20.36      | 20.36        | 67.47 | 13.70             |
| LLaMA2-7B-Chat Inline Quotation (sentence)              | 76.36  | 31.75 | 56.63    | 36.05      | 34.66        | 84.93 | 29.44             |
| gpt-3.5-turbo-0125 ALCE (sentence)                      | 40.89  | 30.06 | 78.39    | 45.92      | 45.92        | 64.82 | 29.77             |
| gpt-3.5-turbo-0125 Inline Quotation (sentence)          | 81.30  | 34.97 | 74.69    | 49.93      | 49.97        | 84.72 | 42.30             |
| LLaMA2-7B-Chat ALCE (subsentence)                       | 65.49  | 29.02 | 48.43    | 5.66       | 5.66         | 93.67 | 5.30              |
| LLaMA2-7B-Chat Inline Quotation (subsentence)           | 78.99  | 30.93 | 53.04    | 10.66      | 24.72        | 91.14 | 22.53             |
| gpt-3.5-turbo-0125 ALCE (subsentence)                   | 45.15  | 33.42 | 77.96    | 19.42      | 19.42        | 80.23 | 15.58             |
| gpt-3.5-turbo-0125 Inline Quotation (subsentence)       | 85.34  | 34.66 | 74.07    | 13.32      | 40.46        | 89.00 | 36.01             |

Note: Here,  $corr (quote) * gran$  evaluates citation the overall citation quality respectively using the product of quote-level citation correctness and granularity; *coarse-paragraph*: combining two or three documents to generate a coarse document; *default-paragraph*: original document in the dataset; *sentence*: splitting documents into sentences; *subsentence*: splitting documents into subsentences.

This ablation study focuses on evaluating the impact of interference documents on answer accuracy and citation appropriateness. Our observations indicate that increased interference from documents leads to reduced answer accuracy. However, our methods demonstrate advantages in citation appropriateness to some extent with the number of interference documents increasing. Interestingly, the citation correctness of the ALCE method does not show an obvious change with the increasing number of interference documents. This can be explained as the model creating incorrect answers based on irrelevant documents, affecting the factualness of the answers. The ablation study of appropriateness highlights the significance of improving citation appropriateness, particularly in scenarios where retrieved documents lack sufficient supporting factual knowledge. This emphasizes the importance of developing methods that can maintain citation appropriateness even when the retrieved documents contain a certain percentage of interference, which is common in real-world scenarios.

### 5.2.2 Granularity Analysis

Moreover, the granularity of reference documents was varied between *default*, *coarse-grained*, and *fine-grained* settings. Here, *coarse-paragraph* means combining two or three documents to generate a coarse document; *default-paragraph* means original document in the dataset; *sentence* means splitting documents into sentences; *subsentence* means further splitting documents into subsentences. In the granularity analysis, we investigated the impact of document granularity on citation correctness and granularity. Our experimental results reveal that coarser document granularity leads to improved citation correctness but reduces citation granularity. Conversely, adopting sentence-level document splitting significantly enhances citation granularity and achieves the best overall citation quality. However, we observed that further splitting documents into subsentence-level increases granularity but sacrifices citation correctness, resulting in a lower overall citation quality. Furthermore, over-fined document granularity is found to confuse LLMs with document indices, leading to reduced document citation correctness. These findings high-light the importance of balancing correctness and granularity for improving citation quality, thus emphasizing the need for careful consideration of document granularity levels to achieve optimal citation outcomes.

## 6. CONCLUSION

In this paper, we introduce a new citation task, AWeCita, aimed at improving LLMs' citation quality in the domain of QA by focusing precisely on citing appropriately in a well-grained way. We also propose two evaluation metrics, citation appropriateness, and citation granularity, to assess LLMs' performance on the AWeCita task. Additionally, we employ various methods to generate citations and conduct experiments using diverse models on different datasets.

Through extensive experimentation, we have gained valuable insights. Notably, we observed that appropriateness positively impacts answer accuracy, particularly when the retrieved documents contain interference and lack factual information. Furthermore, enhancing granularity contributes to improved

citation quality. These findings highlight the significance of employing appropriate and well-grained citation strategies to enhance the overall quality and reliability of question answering systems.

### **7. LIMITATIONS**

Our study has several limitations that should be noted. (1) It relies on an NLI model to assess appropriateness, correctness, and granularity in determining the support of a set of citations for an answer, which potentially introduces evaluation bias. (2) While the metric of citation appropriateness assesses whether an answer relies on any of the retrieved reference documents, it can only filter out cases where the answer is unrelated to the reference documents. It cannot filter out situations where the model generates hallucinatory responses based on interfering documents. Therefore, this metric may not effectively cover all possible situations. (3) Our experiments use a limited number of models. They focus primarily on LLaMA models and a GPT model, which may not fully reflect the performance of all models. This might lead to biased conclusions about citation appropriateness and granularity across different model architectures. (4) The experimental settings lack diversity, as they do not vary the shot count, total number of documents, etc. This limitation affects the generalizability and robustness of the findings. (5) The experiments are limited by the suitability of datasets. The use of ASQA and ELI5 datasets may not adequately capture the appropriateness of citations in generating answers. This limitation restricts the ability to fully explore the impact of citation appropriateness and granularity in real-world question answering tasks. Overall, improving the AWeCita task requires deeper exploration and refinement of evaluation methods, metrics, experiments, and datasets to enhance the reliability and effectiveness of citation evaluation. We have listed these limitations and directions above for future in-depth study.

### **8. DISCUSSION**

To address the limitations of our study, several potential directions can be considered. (1) To mitigate the reliance on an NLI model for evaluation, incorporating GPT-4 or human annotations could provide complementary perspectives and more accurate judgments. (2) To enhance the citation appropriateness metric, integrating answer accuracy into this metric would help ensure both appropriate citations and faithful answers. (3) Future research should include a broader range of models beyond LLaMA and GPT to avoid biased conclusions and reflect diverse architectures. (4) Experimental settings should be varied by adjusting the number of shots, the number of documents, and other parameters to enhance robustness and generalizability. (5) Additionally, expanding the dataset scope beyond ASQA and ELI5 to more complex datasets may better capture citation appropriateness in different contexts.

### **ACKNOWLEDGMENTS**

This work is supported by the National Natural Science Foundation of China (No. 62176257) and the Youth Innovation Promotion Association CAS.

## **AUTHOR CONTRIBUTIONS**

Suifeng Zhao: Performed the research; Wrote the manuscript.

Tong Zhou: Performed the research; Revised the manuscript.

Zhuoran Jin: Proposed the research problems.

Hongbang Yuan: Took part in proposing the research problems.

Yubo Chen: Proposed the research problems.

Kang Liu: Designed the research framework; Revised the manuscript.

Sujian Li: Designed the research framework; Revised the manuscript.

## **REFERENCES**

- [1] Sawicki, J., Ganzha, M., Paprzycki, M., et al.: The State of the Art of Natural Language Processing—A Systematic Automated Review of NLP Literature Using NLP Techniques. *Data Intelligence*, 5(3), 707–749 (2023)
- [2] Schimanski, T., Ni, J.W., Kraus, M., et al.: Towards faithful and robust LLM specialists for evidence-based question-answering. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, pp. 1913–1931 (2024)
- [3] Li, L., Zhang, H., Li, C., et al.: Evaluation on ChatGPT for Chinese Language Understanding. *Data Intelligence*, 5(4), 885–903 (2023)
- [4] Duan, H., Yang, Y., and Kar, Y.T., et al.: Do LLMs know about hallucination? An empirical investigation of LLM’s hidden states. Preprint. arXiv: 2403.10849 (2024)
- [5] Jiang, X.L., Tian, Y.X., Hua, F.R. et al.: A survey on Large Language Model hallucination via a creativity perspective. Preprint. arXiv:2402.06647 (2024)
- [6] Wei, J., Yao, Y., Ton, J.F., et al.: Measuring and reducing LLM hallucination without gold-standard answers via expertise-weighting. Preprint. arXiv:2402.10412 (2024)
- [7] Li, W., Li, J., Ma, W., and Liu, Y., et al.: Citation-enhanced generation for Llm-based chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, pp. 1451–1466 (2024)
- [8] Zhang, Y., Wang, Y., Wang, K., et al.: When Large Language Models meet citation: A survey. Preprint. arXiv:2309.09727 (2023)
- [9] Muther, R., Smith, D.A. : Citations as queries: Source attribution using language models as rerankers. Preprint. arXiv:2306.17322 (2023)
- [10] Gao, T., Yen, H., Yu, J., and Chen, D., et al.: Enabling large language models to generate text with citations. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, Singapore. pp. 6465–6488 (2023)
- [11] Asai, A., Zhong, Z., Chen, D., et al.: Reliable, adaptable, and attributable language models with retrieval. Preprint, arXiv:2403.03187 (2024)
- [12] Yang, Y., Chern, E., Qiu, X., et al.: Alignment for honesty. Preprint, arXiv:2312.07000 (2023)

- [13] Menick, J., Trebacz, M., Mikulik, W., et al.: Teaching language models to support answers with verified quotes. Preprint, arXiv:2203.11147 (2022)
- [14] Stelmakh, I., Luan, Y., Dhingra, B., et al.: ASQA: factoid questions meet long-form answers. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, pp. 8273–8288 (2022)
- [15] Fan, A., Jernite, Y., Perez, E., et al.: ELI5: Long Form Question Answering. In: Proceedings of the 57<sup>th</sup> Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy (Volume 1: Long Papers) pp. 3558–3567 (2019)
- [16] Liu, N., Zhang, T., Liang, P., et al.: Evaluating verifiability in generative search engines. In: Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, pp. 7001–7025 (2023)
- [17] Funkquist, M., Kuznetsov, I., Hou, Y., et al.: Citebench: A benchmark for scientific citation text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, pp. 7337–7353 (2023)
- [18] Sun, H., Cai, H., Wang, B., et al.: Towards verifiable text generation with evolving memory and self-reflection. Preprint, arXiv:2312.09075 (2023)
- [19] Ye, X., Sun, R., Arik, Ö, S., et al.: Effective Large Language Model adaptation for improved grounding and citation generation. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, pp. 6237–6251 (2024)
- [20] Yue, X., Wang, B., Chen, Z., et al.: Automatic evaluation of attribution by Large Language Models. In: Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, pp. 4615–4635 (2023)
- [21] Li, X., Cao, Y., Pan, L., et al.: Towards verifiable generation: A benchmark for knowledge-aware language model attribution. In: Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, pp. 493–516 (2024)
- [22] Huang, J., Chang, K., et al.: Citation: A key to building responsible and accountable Large Language Models. In: Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, pp. 464–473 (2024)
- [23] Wang, Y., Xu, X., Xie, R., et al.: Generative retrieval with large language models. Available at: <https://openreview.net/forum?id=ZdjKRbtrth> (2024) Accessed 22 May 2020
- [24] Honovich, O., Aharoni, R., Herzig, J., et al.: TRUE: re-evaluating factual consistency evaluation. In: Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, DialDoc@ACL 2022, Dublin, Ireland, pp. 161–175 (2022)
- [25] Touvron, H., Martin, L., Stone, K., et al.: Llama 2: Open foundation and fine-tuned chat models. Preprint, arXiv:2307.09288 (2023)

## **AUTHOR BIOGRAPHY**

**Suifeng Zhao** is now pursuing a bachelor's degree in Beijing University of Posts and Telecommunications.

Email: zhaosuifeng@bupt.edu.cn; ORCID: 0009-0004-9859-5782

**Tong Zhou** is currently an engineer at the Institute of Automation, Chinese Academy of Sciences.

Email: tong.zhou@ia.ac.cn; ORCID: 0000-0002-0159-7414

**Zhuoran Jin** is now pursuing a Ph.D. degree in Institute of Automation, Chinese Academy of Sciences.

Email: zhuoran.jin@nlpr.ia.ac.cn; ORCID: 0000-0002-1667-8145

**Hongbang Yuan** is now pursuing a M.S. degree in Institute of Automation, Chinese Academy of Sciences.

Email: hongbang.yuan@nlpr.ia.ac.cn; ORCID: 0009-0003-1178-9756

**Yubo Chen** is an associate professor of Institute of Automation, Chinese Academy of Sciences.

Email: yubo.chen@nlpr.ia.ac.cn; ORCID: 0000-0002-5485-9916

**Kang Liu** is a professor of Institute of Automation, Chinese Academy of Sciences.

Email: kliu@nlpr.ia.ac.cn; ORCID: 0000-0002-6083-8433

**Sujian Li** is a professor in the School of Computer Science, Peking University.

Email: lisujian@pku.edu.cn; ORCID: 0000-0001-7493-0786

APPENDICES

Table 5, table 6, and table 7 demonstrates prompts and output examples for Link Quotation, Inline Quotation, and Post Quotation methods.

Table 5. Link Quotation Example.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Instruction:</b> Generate an answer and a reference span for the given question. For the Answer: Write an accurate, engaging, and concise answer for the given question using the provided search results as a reference (some of which might be irrelevant), and cite them properly. Use an unbiased, journalistic tone. Only cite for the factual claims that the search results support. In the answer, if you need to cite, mark sentences that cite relevant documents with a reference number starting from 1, 2, 3, and so on, incrementing for each subsequent citation. Optionally cite relevant documents and cite at most two documents in each sentence. The situation of citing 123 after one sentence is not allowed. Do not overcite, if multiple documents support the sentence, only cite a minimum sufficient subset of the documents and cite at most two. For the reference span: For sentences marked with reference numbers in the answer, list the corresponding document number, along with the exact citation span.</p> <p>Document [1] (Title: Cherrapunji): Cherrapunji Cherrapunji; (with the native name Sohra being more commonly used, and can also be spelled Cherrapunjee or Cherrapunji) is a subdivisional town in the East Khasi Hills district in the Indian state of Meghalaya. It is the traditional capital of aNongkhlaw ĩ imaw hima (Khasi tribal chieftainship constituting a petty state), both known as Sohra or Churra. Cherrapunji has often been credited as being the wettest place on Earth, but for now nearby Mawsynram currently holds that distinction. Cherrapunji still holds the all-time record for the most rainfall in a calendar month for July 1861 and most rain in a year from August 1860 to July 1861, however: it received in</p> <p>Document [2] (Title: Cherrapunji): Radio relay station known as Akashvani Cherrapunji. It broadcasts on FM fre- quencies. Cherrapunji Cherrapunji; (with the native name Sohra being more commonly used, and can also be spelled Cherrapunjee or Cherrapunji) is a subdivisional town in the East Khasi Hills district in the Indian state of Meghalaya. It is the traditional capital of aNongkhlaw hima (Khasi tribal chieftainship constituting a petty state), both known as Sohra or Churra. Cherrapunji has often been credited as being the wettest place on Earth, but for now nearby Mawsynram currently holds that distinction. Cherrapunji still holds the all-time record for the most rainfall"</p> <p>Document [3]...</p> <p>Document [4]...</p> <p>Document [5]...</p> <p><b>Answer:</b></p> <p><b>Answer:</b></p> <p>Several places on Earth claim to be the most rainy, such as Lloro, Colombia, which reported an average annual rainfall of 12,717 mm between 1952 and 1989, and Lopez de Micay, Colombia, which reported an annual 12,892 mm between 1960 and 2012 {1}. However, the official record is held by Mawsynram, India with an average annual rainfall of 11,872 mm {2}, although nearby town Sohra, India, also known as Cherrapunji, holds the record for most rain in a calendar month for July 1861 and most rain in a year from August 1860 to July 1861 {3}.</p> <p><b>Reference:</b></p> <p>1. Document [3]: but that claim is disputed by Lloro, Colombia, which reported an average yearly rainfall of 12,717 mm between 1952 and 1989 and Lopez deMicay, also in Colombia, which reported an annual 12,892 mm per year between 1960 and 2012.</p> <p>2. Document [3]: Mawsynram receives one of the highest rainfalls in India. It is reportedly the wettest place on Earth, with an average annual rainfall of 11,872 mm,</p> <p>3. Document [1]: Cherrapunji still holds the all-time record for the most rainfall in a calendar month for July 1861 and most rain in a year from August 1860 to July 1861</p> |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

**Table 6.** Inline Quotation Example.

---

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Instruction:</b> Craft a detailed and engaging response to the question using excerpts from provided documents. To ensure accuracy and relevance, embed citations directly into your answer by using latex footnote format footnoteFrom document [document id]: continuous text fragment in this document literally, quoting the text frag- ments verbatim within brackets. Cite only when stating facts supported by the documents, using a maximum of two references per sentence. When multiple documents corroborate a sentence, choose only the essential ones for cita- tion. Incorporate personal insights or connections to bridge cited information, enhancing the narrative flow without compromising factual integrity. Avoid excessive citation; aim for a balanced and insightful reply.</p> <p>Document [1]:. . .</p> <p>Document [2]:. . .</p> <p>Document [3]:. . .</p> <p>Document [4]:. . .</p> <p>Document [5]:. . .</p> <p><b>Question:</b> When did the us break away from england?</p> <p><b>Answer:</b></p> <p>The United States took the first step towards gaining independence from Great Britain when it declared independence from Great Britain on July 2, 1776 (although the event is now commemorated on July 4, 1776, the date when the Declaration of Independence was officially adopted by Congress) \\footnote{From document [2]: "The United States of America declared independence from Great Britain on July 2, 1776"} \\footnote{From document [1]: "On July 4th, the Continental Congress formally adopted the Declaration of Independence"}. The Treaty of Paris was later signed on September 3, 1783, formally separating the United States from the British Empire \\footnote{From document [3]: "The Treaty of Paris was signed September 3, 1783, formally ending the conflict and confirming the new nation's complete separation from the British Empire."}.</p> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

---

**Table 7.** Post Quotation Example.

---

**Instruction:** Please list detailed quotes from the documents in the order of the reference numbers cited in the Answer. Ensure that each quote provided directly supports the related sentence in the answer. Each quote must correspond to a reference mark [Document ID] mentioned in the Answer, clearly linking the evidence to its respective assertion. The format for each quote should be: From Document [document ID]: a continuous text fragment exactly as it appears in the document literally.

There are some examples. Complete the last Quote according to example format.

**Question:** Which is the most rainy place on earth?

**Response:** 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 [5]. However, the official record is held by Mawsynram, India with an average annual rainfall of 11,872 mm [5], although nearby town Sohra, India, also known as Cherrapunji, holds the record for most rain in a calendar month for July 1861 and most rain in a year from August 1860 to July 1861 [4].

Document [1]:. . .

Document [2]:. . .

Document [3]:. . .

Document [4]:. . .

Document [5]:. . .

**Quote:**

From Document [5]: but that claim is disputed by Lloro, Colombia, which reported an average yearly rainfall of 12,717 mm between 1952 and 1989 and Lopez deMicay, also in Colombia, which reported an annual 12,892 mm per year between 1960 and 2012.

From Document [5]: Mawsynram receives one of the highest rainfalls in India. It is reportedly the wettest place on Earth, with an average annual rainfall of 11,872 mm

From Document [4]: Cherrapunji still holds the all-time record for the most rainfall in a calendar month for July 1861 and most rain in a year from August 1860 to July 1861

---