CiteVerifier: How Good Are Citation Verifiers and How to Use Them?

Konrad Wojtasik[1][0000−0002−5715−5201], Tymoteusz Dolega[1][0009−0001−8376−1299], and Maciej Piasecki[1][0000−0003−1503−0993]

Wrocław University of Science and Technology, Poland

Abstract. Large language models (LLMs) have become powerful tools for understanding documents and answering questions (QA). The grounding of these answers consistently in facts in the given documents may be achieved by citing them in the generated responses. Several approaches to Retrieval Augmented Generation (RAG) have been proposed that incorporate citation to relevant documents to enhance correctness and verifiability. However, evaluation if the document is cited accurately, relies heavily on large generative models for Natural Language Inference. In this work, we evaluate various models in different evaluation schemes for the citation verification task to provide insights into how these models perform and in which evaluation schemes they excel. Our findings show that the TRUE T5 model performs well in verifying the completeness of citations, but struggles when only partial information is available. We also demonstrate that general LLMs can perform citation verification effectively, although the results in citation addition on an already generated answer as post-processing are still suboptimal. We argue that it is important to be mindful of how citation verifiers are used and understand their strengths and limitations. Furthermore, we trained a small and lightweight model, *CiteVerifier*, which performs exceptionally well despite being magnitudes smaller than other models, making it an ideal solution for low-resource settings.

Keywords: Retrieval Augmented Generation · Natural Language Processing

1 Introduction

The most convenient way of question answering today is to use Large Language Models (LLMs), which mostly perform well in QA and generate coherent responses [9]. However, while LLM answers may seem plausible, they are susceptible to hallucinations and false claims [8]. Ensuring that the generated information is accurate poses a challenge, especially in dynamic domains. Retrieval-Augmented Generation (RAG), which pulls in answers from unstructured sources, is a path to make LLM-based QA more reliable [6]. Adding citations to the original sources can further improve accuracy, trustfulness and comprehension of the answers, as the citations facilitate verification of the information by its source [5]. Citations can be added in a post-generation process

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_19

or generated along with the answer from the retrieved documents [19]. Citation correction verification typically involves the use of fine-tuned Natural Language Inference (NLI) models or LLMs as evaluators [16]. In this work, we evaluate various models across different schemes to gain a comprehensive understanding of their performance in the citation verification task. Our goal is to identify where these models tend to make mistakes and analyze their specific characteristics in relation to the schemes in which they are applied. This analysis is crucial for improving the quality of citation evaluation metrics and enhancing post-generation answer validation in applied systems, where correctness is essential.

We observed that some evaluation models are sensitive to the completeness of citations, expecting all necessary documents to be cited in answers. If the information provided in documents is incomplete, they tend to return false entailment, even if the claim is partially supported. This is particularly visible for the widely used NLI model, TRUE [7]. However, some models can effectively detect whether the claims in the generated answer are partially supported by the cited document. Using these models, there is no guarantee that the information is complete and there might always be some extra claim in the answer that requires verification.

In addition, we evaluated *post-citing task*, in which the correct citations must be selected for each sentence in a generated answer from a list of candidate documents provided. We found that most evaluation methods demonstrate poor overall accuracy, despite performing well in evaluating individual citations.

Furthermore, we prepared a bilingual dataset in both English and Polish. To our knowledge, this is the first work to evaluate the performance of citation verification models for the Polish language.

The main contributions of our work can be summarized as follows.

- We prepared a bilingual dataset based on the HotpotQA [18] dataset in both English and Polish, specifically tailored for evaluating citation verification models. This dataset is the first of its kind to support citation verification tasks in Polish, providing a new resource for future research. We also include a training dataset used for training our *CiteVerifier* model.
- We evaluated a range of citation verification models, including fine-tuned NLI models, LLMs and trained our own lightweight *CiteVerifier* model which achieved high performance, while being multiple times smaller than the tested models. Our analysis not only presents a detailed comparison of their performance but also highlights their pros and cons.
- We demonstrated that post-citation is a challenging task even for LLMs, as evidenced by their low overall accuracy. Our findings suggest that while models can often verify individual citations, their ability to select all necessary citations from a list of candidates is still limited, pointing to the need for further advancements in this area.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_19

2 Related Work

Recently, various approaches to training models capable of citing sources in their answers have been proposed. GopherCite [12] uses an LLM trained with reinforcement learning from human feedback (RLHF) and supervised learning to generate answers supported by evidence retrieved via Google search. The model was evaluated by paid contractors on a manageable test sample.

ALCE [5] represents the first systematic attempt to develop fundamental methods enabling LLMs to generate citations in QA tasks. It also introduces metrics for evaluating citation quality, highlighting that there is still room for improvement in citation generation. The authors employ the TRUE NLI model [7] to automatically verify whether the cited passage entails the model-generated response and conduct experiments demonstrating a strong correlation with human evaluation. Based on the NLI relations, they propose a method to calculate the citation recall and precision. Citation recall aligns with our Merged Evaluation scheme, introduced in 3.4, where concatenated citations must collectively support the statement. A key drawback of this metric is that a citation may be deemed irrelevant, even if it provides partial support for the claim. Citation precision, on the other hand, requires the concatenated cited documents to fully entail the statement, that happens only when citation recall equals 1. The relevance of a citation is then assessed by determining whether an individual citation fails to support the claim and removing this citation from the concatenated documents still allows the claim to be supported, the citation is classified as irrelevant. We propose Separate Evaluation to eliminate the requirement for complete entailment of the statement by the entire set of cited documents. This approach allows for the identification of partial support for claims, but can not guarantee completeness.

The Self-RAG framework [1] introduces retrieval and critique tokens, enabling flexible retrieval-augmented generation with citations by incorporating feedback mechanisms for improved output selection. Efficient Citer [17] trains a 3B Flan-T5 and a 13B LLaMA models on a dataset based on MS MARCO [14], where the answers were generated by ChatGPT with incorporated citations. Both Self-RAG and Efficient Citer incorporated ALCE [5] in their evaluation.

Similarly, the AGREE framework [19] employs the TRUE NLI model to notate citations in the answers. The framework fine-tunes an LLM to generate responses grounded in citations and capable of identifying unsupported claims, addressing these claims during test-time adaptation (TTA) by iteratively searching for additional evidence. Another related approach is proposed in [4], where the Attributable to Identified Sources (AIS) metric is introduced. This metric, implemented using the NLI-based TRUE model, evaluates the accuracy of the attribution by determining whether each sentence of the answer is entailed by evidence from a retrieved set.

Other methods explore different models and strategies for citation evaluation and generation. CEG [11] proposes using LLMs (GPT-3.5-Turbo and GPT-4) as NLI evaluators to add citations to generated answers in a post-hoc manner. This

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_19

approach assesses the factuality of claim-document pairs, regenerating answers to correct identified errors.

3 Methodology

3.1 Data Preparation

For our experiments, we used a subset of HotpotQA [18]. For each question, an answer with citations was generated using the Command R+ model[1] , an LLM specialized for RAG. From them, 1000 randomly selected answers were chosen for further evaluation. The citations were assessed using the TRUE model [7], which helped to filter the answers. After filtering, 399 responses remained and of these, 100 correct answers were manually verified for the experiments. For each correct answer, a related incorrect one was manually created by significantly altering the core information to make it inconsistent with the corresponding documents, but still plausible.

For each sample instance, 10 documents were provided, including 2 documents relevant for the answer. Each answer was divided into individual sentences, and each was assigned the appropriate documents as citations to them. It is important to note that while there were always 2 relevant documents for each answer, each sentence in it is assigned 1-2 citations.

The dataset was subsequently automatically translated into Polish using *GPT-4o-mini* [15] via the OpenAI API. All translated answers were manually verified and corrected.

3.2 CiteVerifier Model and Training Data

Our lightweight *CiteVerifier* model, proposed as a strong baseline, is based on the HerBERT-large [13] model for Polish, initialized from XLM-Roberta-large [2] and next fine-tuned. Thus, we could train it directly on English data while achieving strong performance on both English and Polish evaluation datasets.

The training dataset of *CiteVerifier* has been derived from HotpotQA, and consists of 10,000 correct document-sentence pairs, i.e. the documents are proper citations for the sentences. In addition, there are 10,000 incorrect document-sentence pairs, where the documents are not relevant to the sentences. Finally, the dataset also includes 10,000 modified sentences with the crucial information altered, that breaks the citation link to the original documents. These modifications to the sentences were generated with the Llama 3.1 70B [3] and Command R+ models and after manual verification we noticed that the Llama examples were more focused on details and better suited as training examples.

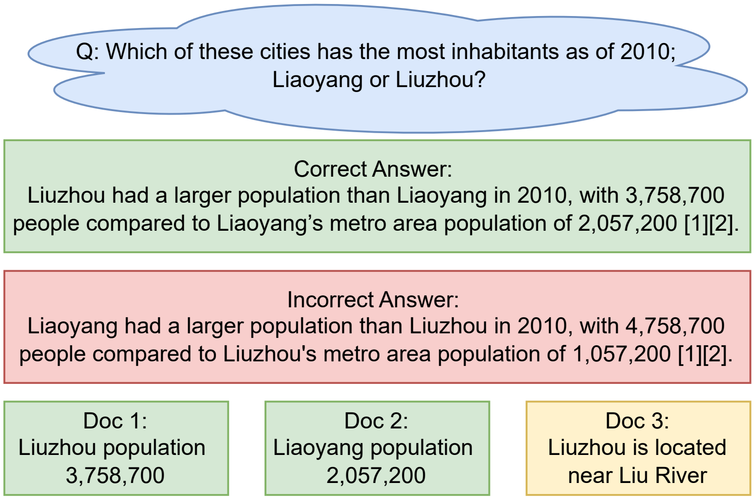
[1] https://huggingface.co/CohereForAI/c4ai-command-r-plus

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_19

3.3 Evaluated Models

We evaluated a wide range of models, all set up using the vLLM framework [10]. Command R+ model was run with FP8 precision to further minimize GPU memory usage, while the other models were run with FP16 precision. We evaluated 7 different models: Llama 3 8B [2], Llama 3.1 70B[3], Command R+, Bielik v2.2 11B[4], TRUE T5[5], mDeberta[6] and the GPT-4o model through OpenAI API (version 2024-08-06).

3.4 Evaluation Settings



Q: Which of these cities has the most inhabitants as of 2010; Liaoyang or Liuzhou?

Correct Answer:
Liuzhou had a larger population than Liaoyang in 2010, with 3,758,700 people compared to Liaoyang's metro area population of 2,057,200 [1][2].

Incorrect Answer:
Liaoyang had a larger population than Liuzhou in 2010, with 4,758,700 people compared to Liuzhou's metro area population of 1,057,200 [1][2].

Doc 1:
Liuzhou population
3,758,700

Doc 2:
Liaoyang population
2,057,200

Doc 3:
Liuzhou is located
near Liu River

Fig. 1. *Correct Answer* includes accurate citations of the relevant documents, and the statements align with the information provided in those documents. *Incorrect Answer*, contains statements that contradict the information from the cited documents. Document 3 presents an *Incorrect Document*, that can be added to correct documents or replace a correct document in Merged Evaluation scheme.

The evaluation was performed according to 3 main evaluation schemes.

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[3] https://huggingface.co/meta-llama/Meta-Llama-3.1-70B-Instruct

[4] https://huggingface.co/speakleash/Bielik-11B-v2.2-Instruct

[5] https://huggingface.co/google/t5_xxl_true_nli_mixture

[6] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli

Merged Evaluation In this scheme, cited documents are collectively evaluated to determine whether they collectively support the answer sentence, or it can be inferred from them. This evaluation is particularly well-suited for NLI models, which are designed to determine whether a premise completely entails a hypothesis.

All necessary information for validating the entailment must be contained within the documents considered as a premise. A model evaluates the documents concatenated together to determine, if they collectively provide sufficient evidence for the given answer. This approach requires a model to integrate information from multiple sources and make a judgment based on the collective evidence, rather than analyzing each document in isolation. For this scheme, we propose five experiment settings:

Correct Answer – a model is tested using only accurate documents that contain the correct information and relevant sentences. The aim here is to see how well the model can understand and respond to questions based on factual, reliable sources without any distractions, e.g., in Fig. 1, this is a case where *Correct Answer* is checked with Doc 1 and Doc 2 concatenated.

Incorrect Answer – a model is tested with answers including crucial information intentionally modified, e.g., key facts are changed, causing the answer to look correct but being false, e.g., in Fig. 1 this is shown as *Incorrect Answer* checked with concatenated Doc 1 and Doc 2.

Replaced Document – one of the correct documents is replaced with a document chosen from a set of 10 documents that are on a topic similar to the question but not related to the answer. The information for the entailment is incomplete, but partly relevant, with additional distraction from the replaced document, in Fig 1 it equals to comparing *Correct Answer* with Doc 1 and Doc 3, instead of Doc 2.

Added Document – a random document or two documents were added to the two correct documents relevant to the answer, e.g. in Fig 1 it means to compare *Correct Answer* with Doc 1, Doc 2 and Doc 3 concatenated together. So, there is some extra, non-related information present, and we expect to observe some influence on the predictions of the model.

Removed Document – one of the correct documents is removed, so the information for entailment is incomplete, but there is no additional distraction, e.g., in Fig. 1, this is a case where *Correct Answer* is compared with only Doc 1.

Separate Evaluation Each document is individually evaluated to determine, if the model can properly assess, whether a document partially supports the given answer. The goal is to evaluate the model's ability to analyze individual documents as offering relevant, but possibly incomplete evidence for the answers. This scheme is not suitable for NLI models, as the entitlement relation is not satisfied with incomplete information.

Three settings are proposed:

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_19

Correct Answer – each document that contains the correct relevant information is paired with the corresponding sentence of the answer, e.g., in Fig. 1, *Correct Answer* is evaluated individually with Doc 1 and Doc 2.
Incorrect Answer – the answer has been intentionally modified so that it is no longer supported by the document. The goal is to assess, if the model is able to spot inconsistencies or false statements when presented with misleading information, e.g., to evaluate *Incorrect Answer* with Doc 1 or Doc 2.
Incorrect Document – the answer itself is correct, but the document provided as a citation is not relevant. The incorrect documents are randomly selected from the distractor documents of the original HotpotQA. This experiment aims to test whether the model can discern that the document fails to support the correct answer, even though the answer is valid. It checks the model's ability to reject irrelevant or inaccurate citations.

Citation In this scheme, the core task is to choose the correct documents to be cited from among the 10 documents provided. This scheme closely resembles the RAG use case, in which, after answer generation, appropriate citations are assigned to each sentence of the response. The goal is to identify which documents provide the necessary evidence to support the information in the generated answer.

This task differs from a simple citation verification process. In citation verification, the focus is primarily on confirming whether the citations already present are accurate and directly related to the answer. However, this task is more challenging: multiple documents must be evaluated for each answer's sentence, and relevant documents recognized. Additionally, there is higher likelihood that many documents are not relevant, so the task involves filtering through potentially distracting or unrelated content. The models were evaluated in three tasks.
Separate Document – each document is presented individually, in similar way to *Separate Evaluation*, for each document, the model is run to determine whether it should be cited for a particular answer sentence.This method evaluates the model's ability to assess the relevance of each individual document for a given sentence.
Merged Documents – initially, all documents are presented together. Subsequently, one document is removed at a time. If this results in false entailment relation, then the removed document is considered crucial and relevant. This approach is particularly beneficial for evaluating NLI models that require a complete set of premises to establish entailment with a hypothesis. By observing how the absence of individual documents affects the entailment relationship, this experiment assesses the model's reliance on the completeness of information for accurate reasoning. A similar approach was proposed in ALCE.
List of Documents – an LLM chooses the right documents that should be cited for a given sentence. A list of documents, labeled with the corresponding numbers, is presented, all documents at the same time. Due to the generative nature of LLMs and the ability to process large contexts, this seems to be an adequate way to set up this task and let LLM generate the right document indexes.

Additionally, this approach requires only a single model inference rather than multiple inferences to evaluate each document individually, potentially saving significant computational resources.

4 Experiments

4.1 Merged Evaluation

Merged Evaluation English Results

Model Type	Prompt	**Accuracy**	**Acc True**	**Acc False**	**F1**
gpt-4o	0shot_en	0.8771	0.9065	0.8528	0.8694
T5 TRUE	-	0.8569	0.9228	0.8027	0.8534
Bielik 2 11B	cot_en	0.8404	0.9472	0.7525	0.8427
Llama 3.1 70B	0shot_en	0.8532	0.8374	0.8662	0.8374
Command R+	cot_en	0.8239	0.9228	0.7425	0.8255
Llama 3 8B	0shot_en	0.7376	0.9715	0.5452	0.7697
mDeBERTa	-	0.6991	0.6341	0.7525	0.6555

Merged Evaluation Polish Results

Model Type	Prompt	**Accuracy**	**Acc True**	**Acc False**	**F1**
Bielik 2 11B	cot_pl	0.8275	0.9593	0.7191	0.8339
Command R+	0shot_en-2	0.8110	0.9187	0.7224	0.8144
gpt-4o	0shot_pl	0.8385	0.7764	0.8896	0.8128
Llama 3.1 70B	ceg1	0.7908	0.8252	0.7625	0.7808
Llama 3 8B	ceg1	0.7009	0.9593	0.4883	0.7433
T5 TRUE	-	0.7578	0.6098	0.8796	0.6944
mDeBERTa	-	0.6606	0.6260	0.6890	0.6247

Table 1. Merged scheme evaluation. The documents texts are concatenated and checked with the answer sentence as a hypothesis. In this scheme, it checks if the model is sensitive to the completeness of the information in the provided documents.

In *Merged Evaluation* scheme, as shown in Tab. 1, gpt-4o model achieved the best results for English, which is not surprising since this is a well-trained commercial model. T5 TRUE model, the second place, is recently frequently used in citation verification and applied in evaluation of LLM generated answers with citations. The results confirm that it is very good at assessing the completeness of the citations. Unfortunately, it does not perform well on the Polish evaluation data set, as it has been trained for English. For the Polish language, the best results were achieved by Bielik 2 11B, demonstrating its strong ability to perform the NLI task effectively, with surprisingly high performance also for the English language. The small multilingual NLI model, mDeBERTa, despite its multilingual capabilities, obtained the lowest scores for both languages, suggesting that larger models can better capture the completeness of information.

4.2 Separate Evaluation

Separate Evaluation English Results

Model Type	Prompt	**Accuracy**	**Acc True**	**Acc False**	**F1**
Command R+	ceg1	0.8586	0.9633	0.8084	0.8155
Llama 3 8B	ceg1	0.8542	0.8807	0.8414	0.7967
CiteVerifier	-	0.8423	0.8028	0.8612	0.7675
Command R+	0shot_en	0.8661	0.6789	0.9559	0.7668
Bielik 2 11B	ceg1	0.8571	0.6651	0.9493	0.7513
Llama3.1 70B	ceg1	0.8229	0.4908	0.9824	0.6426
T5 TRUE	-	0.7857	0.3899	0.9758	0.5414

Separate Evaluation Polish Results

Model Type	Prompt	**Accuracy**	**Acc True**	**Acc False**	**F1**
Llama 3 8B	0shot_pl	0.8348	0.8073	0.8480	0.7603
Command R+	ceg1	0.8036	0.9587	0.7291	0.7600
Bielik 2 11B	ceg1	0.8571	0.6514	0.9559	0.7474
CiteVerifier	-	0.8289	0.7431	0.8700	0.7380
gpt-4o	ceg1	0.8527	0.5642	0.9912	0.7130
Llama 3.1 70B	0shot-post-pl-4	0.8155	0.5046	0.9648	0.6395
T5 TRUE	-	0.7515	0.2706	0.9824	0.4140

Table 2. Separate scheme evaluation. The documents texts are separately checked with the answer sentence as a hypothesis. It checks if the model is sensitive to partial claims that supports the answer sentence.

Separate Evaluation in Tab. 2 shows that, with appropriate prompting, LLMs can effectively determine whether a given document should be cited in an answer. Notably, Llama 3 8B performs exceptionally well, despite being significantly smaller than Llama 3.1 70B or Command R+.

Our *CiteVerifier* model with only 355 million parameters (based on HerBERT-large) also expresses performance competitive to much larger models. So, *CiteVerifier* can be a viable option in low-resource environments and offers an efficient alternative in citation verification tasks.

Moreover, T5 TRUE performed poorly in this scheme due to the incompleteness of the information in the individual documents. Most of the claims required two documents for complete entailment, that likely impacted its performance, as it performs best with comprehensive document sets, as indicated in the Merged Evaluation Results. Additionally, the performance of Llama 3.1 70B dropped significantly in this scheme, suggesting that the model is also sensitive to the completeness of the information provided in the documents.

4.3 Citation Evaluation

Citation Evaluation in Tab 3 reveals that Command R+ achieves the best performance. Notably, changes in the prompt have significant impact on trade-off between *true* and *false* accuracy, highlighting the model sensitivity to prompt formulation. Prompts with the best performance are provided in our GitHub repository.

Despite this, the overall exact accuracy scores remain relatively modest. Command R+ achieves 68% of exact accuracy, while Llama 3.1 70B scores 62%. These figures signal substantial room for improvement. When considering systems designed to provide accurate citations in a post-generation manner, the target accuracy should approach at least 90%. This highlights a significant gap between the current model capabilities and the requirements of practical applications.

Our *CiteVerifier* performs well in this scheme (especially concerning its size!), defending its position against significantly larger models. It achieves a high F1 score with 80% accuracy on *true* samples, while maintaining *Exact Accuracy* of 58% for English language, and demonstrates its efficiency and effectiveness even with fewer parameters. The results for Polish are not far behind, even with a smaller performance gap between the scores of Command R+ in this language.

Separate Citaion Evaluation English Results

Model Type	Prompt	**Accuracy**	**Acc True**	**Acc False**	**F1**	**Exact**
Command R+	0shot_post_en-4	0.9555	0.8165	0.9688	0.7623	0.6786
Command R+	0shot_en	0.9579	0.6789	0.9846	0.7382	0.6825
CiteVerifier	None	0.9471	0.8028	0.9609	0.7261	0.5833
Bielik 2 11B	0shot_post_en-4	0.9395	0.7936	0.9535	0.6962	0.5476
Llama 3 8B	0shot_post_en-4	0.9395	0.6606	0.9662	0.6560	0.5794
Llama 3.1 70B	0shot_post_en-4	0.9455	0.5183	0.9864	0.6243	0.6230
gpt-4o	0shot_post_en-4	0.9443	0.4174	0.9947	0.5670	0.6190
T5 TRUE	None	0.9403	0.3899	0.9930	0.5329	0.5833

Separate Citation Evaluation Polish Results

Model Type	Prompt	**Accuracy**	**Acc True**	**Acc False**	**F1**	**Exact**
Command R+	0shot_pl	0.9535	0.6789	0.9798	0.7184	0.6389
CiteVerifier	None	0.9467	0.7431	0.9662	0.7090	0.5754
Bielik 2 11B	0shot_post_pl-4	0.9415	0.7569	0.9592	0.6933	0.5238
Command R+	0shot_post_pl-4	0.9175	0.8899	0.9201	0.6532	0.5079
gpt-4o	0shot_post_pl-4	0.9479	0.5183	0.9890	0.6348	0.6270
Llama 3 8B	0shot_pl	0.9139	0.8119	0.9236	0.6221	0.4960
Llama 3.1 70B	0shot_post_pl-4	0.9395	0.5000	0.9816	0.5908	0.6151

Table 3. Separate post-citing evaluation results. The task is to select which documents should be cited from a set of 10 provided documents. In this scheme, each document is presented independently, and the model must decide whether or not it should be cited for the given answer. The goal is to evaluate the model's ability to assess each document in isolation and determine its relevance to the answer.

In *Merged Documents Citation*, Tab. 4, only LLMs with long context length are suitable for it. The results show that Llama 3.1 70B excels in this scheme, effectively performing the NLI task on the provided documents and accurately assessing the completeness of information. This high accuracy is crucial for detecting whether one of the required document is missing. The top *Exact* accuracy scores are comparable to the highest exact scores achieved in *Separate Citation* for both English and Polish.

List of Documents Citation, in a similar way to *Merged Documents Citation*, is suitable only for models with long context lengths. In this scheme, the model has access to all documents simultaneously and must determine which documents should be cited. This approach requires a more sophisticated understanding of the task by the model. As shown in Tab. 5, gpt-4o excels in this scheme, and achieves a significant margin in terms of the *Exact* accuracy in comparison to other LLMs. It deliverers the highest *Exact* accuracy across all schemes: 80% for English and 76% for Polish. This scheme proves to be particularly challenging for other LLMs, whose performance drop significantly in comparison to other evaluation schemes.

Merged Documents Citation Evaluation English Results

Model Type	Prompt	**Accuracy**	**Acc True**	**Acc False**	**F1**	**Exact**
Llama 3.1 70B	0shot_en	0.9511	0.5046	0.9939	0.6433	0.6706
Command R+	0shot_en	0.9399	0.5000	0.9820	0.5924	0.5833
gpt-4o	0shot_en	0.9343	0.5917	0.9671	0.6114	0.6627
Bielik 2 11B	0shot_en	0.9371	0.4954	0.9794	0.5791	0.5913
Llama 3 8B	0shot_en	0.9271	0.3394	0.9833	0.4485	0.5516

Merged Documents Citation Evaluation Polish Results

Model Type	Prompt	**Accuracy**	**Acc True**	**Acc False**	**F1**	**Exact**
gpt-4o	0shot_pl	0.9435	0.5000	0.9860	0.6072	0.6825
Llama 3.1 70B	0shot_pl	0.9439	0.4817	0.9881	0.6000	0.6429
Bielik 2 11B	ceg1	0.9407	0.4725	0.9855	0.5819	0.5992
Command R+	0shot_pl	0.9403	0.4725	0.9851	0.5803	0.5635
Llama 3 8B	0shot_pl	0.9155	0.2982	0.9745	0.3812	0.5159

Table 4. Merged post-citing evaluation results. All documents are presented to the model at once, except for one document, which is withheld. If the model outputs that there is no entailment, this suggests that the missing document was crucial and should be cited. This approach is preferential for models that are sensitive to missing information, but requires larger context length.

List of Documents Citation Evaluation English Results

Model Type	Prompt	**Accuracy**	**Acc True**	**Acc False**	**F1**	**Exact**
gpt-4o	select_en	0.9760	0.8899	0.9842	0.8661	0.8095
Llama 3.1 70B	select_en	0.9391	0.8807	0.9447	0.7164	0.5000
Bielik 2 11B	select_en	0.9067	0.8853	0.9087	0.6236	0.3889
Command R+	select_en	0.8886	0.8119	0.8960	0.5601	0.2778
Llama 3 8B	select_en	0.8886	0.7706	0.8999	0.5472	0.3135

List of Documents Citation Evaluation Polish Results

Model Type	Prompt	**Accuracy**	**Acc True**	**Acc False**	**F1**	**Exact**
gpt-4o	select_pl	0.9688	0.8578	0.9794	0.8274	0.7619
Llama 3.1 70B	select_pl	0.9427	0.7431	0.9618	0.6938	0.4960
Bielik 2 11B	select_pl	0.9167	0.8257	0.9254	0.6338	0.3452
Command R+	select_pl	0.9030	0.8486	0.9083	0.6046	0.3214
Llama 3 8B	select_pl	0.9026	0.6514	0.9267	0.5389	0.3135

Table 5. List of Documents Citation evaluation results scheme, assesses the performance of LLMs in selecting the appropriate documents from a provided list that should be cited in support of the answer.

5 Conclusions

We presented evaluation results across various schemes for citation verification and post-citation tasks. We argue that current evaluation metrics for citation-grounded models need reassessment, as is demonstrated by TRUE T5 performance. While TRUE T5 excels in scenarios with complete citations, it struggles in situations in which documents only partially support a claim. This limitation highlights the need to design new evaluation metrics in which LLMs or models like *CiteVerifier* are applied. We have shown that LLMs can achieve strong performance in different schemes, but their effectiveness depends on prompt design. Their long context length is essential for *Merged Documents Citation* scheme and *List of Documents Citation* scheme, where many documents are presented at the same time. The performance in post-citing tasks reveals a significant gap between commercial models like GPT-4o and open-weight models, which indicates potential need of fine-tuning models to the task.

The *CiteVerifier* model, with only 355 million parameters, appears to be an efficient alternative in low-resource settings, offering competitive performance despite its much smaller (in fact, tiny) size.

6 Limitations

Despite extensive experiments, there are still areas that have not yet been explored. Open LLMs have not been specifically fine-tuned for citation verification tasks, which could lead to significant improvements, if such fine-tuning is applied. Additionally, while we have tested various prompts, there is still potential for further optimization, and some prompts might yet be discovered that could enhance overall performance.

Finally, our evaluation was conducted on a subset of the HotpotQA dataset, which only includes cases with a maximum of two cited documents per sentence. In practice, there may be scenarios that involve multiple documents, suggesting that further testing with more complex citation scenarios could provide additional insights.

Acknowledgements

Financed by: (1) CLARIN ERIC (2024–2026), funded by the Polish Minister of Science (agreement no. 2024/WK/01); (2) CLARIN-PL, the European Regional Development Fund, FENG programme (FENG.02.04-IP.040004/24); (3) statutory funds of the Department of Artificial Intelligence, Wroclaw Tech; (4) the EU project "DARIAH-PL", under investment A2.4.1 of the National Recovery and Resilience Plan. The views expressed are those of the authors and do not necessarily reflect those of the EU or the European Research Executive Agency.

References

1. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-rag: Learning to retrieve, generate, and critique through self-reflection (2023), https://arxiv.org/abs/2310.11511
2. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747
3. Dubey, A., et al.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
4. Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A.T., Fan, Y., Zhao, V., Lao, N., Lee, H., Juan, D.C., Guu, K.: RARR: Researching and revising what language models say, using language models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 16477–16508. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.910, https://aclanthology.org/2023.acl-long.910
5. Gao, T., Yen, H., Yu, J., Chen, D.: Enabling large language models to generate text with citations. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 6465–6488. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.398, https://aclanthology.org/2023.emnlp-main.398
6. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: A survey (2024), https://arxiv.org/abs/2312.10997

7. Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., Matias, Y.: TRUE: Re-evaluating factual consistency evaluation. In: Feng, S., Wan, H., Yuan, C., Yu, H. (eds.) Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering. pp. 161–175. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.dialdoc-1.19, https://aclanthology.org/2022.dialdoc-1.19
8. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. CoRR **abs/2202.03629** (2022), https://arxiv.org/abs/2202.03629
9. Kamalloo, E., Dziri, N., Clarke, C., Rafiei, D.: Evaluating open-domain question answering in the era of large language models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5591–5606. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.307, https://aclanthology.org/2023.acl-long.307
10. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (2023)
11. Li, W., Li, J., Ma, W., Liu, Y.: Citation-enhanced generation for LLM-based chatbots. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1451–1466. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024), https://aclanthology.org/2024.acl-long.79
12. Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G., McAleese, N.: Teaching language models to support answers with verified quotes (2022), https://arxiv.org/abs/2203.11147
13. Mroczkowski, R., Rybak, P., Wróblewska, A., Gawlik, I.: HerBERT: Efficiently pretrained transformer-based language model for Polish. In: Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. pp. 1–10. Association for Computational Linguistics, Kiyv, Ukraine (Apr 2021), https://www.aclweb.org/anthology/2021.bsnlp-1.1
14. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: Ms marco: A human generated machine reading comprehension dataset (November 2016), https://www.microsoft.com/en-us/research/publication/ms-marco-human-generated-machine-reading-comprehension-dataset/
15. OpenAI: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774
16. Shen, J., Zhou, T., Zhao, S., Chen, Y., Liu, K.: Citekit: A modular toolkit for large language model citation generation (2024), https://arxiv.org/abs/2408.04662
17. Tahaei, M., Jafari, A., Rashid, A., Alfonso-Hermelo, D., Bibi, K., Wu, Y., Ghodsi, A., Chen, B., Rezagholizadeh, M.: Efficient citer: Tuning large language models for enhanced answer quality and verification. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024. pp. 4443–4450. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.findings-naacl.277, https://aclanthology.org/2024.findings-naacl.277
18. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of

the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2369–2380. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1259, https://aclanthology.org/D18-1259

19. Ye, X., Sun, R., Arik, S., Pfister, T.: Effective large language model adaptation for improved grounding and citation generation. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 6237–6251. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.naacl-long.346, https://aclanthology.org/2024.naacl-long.346

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_19