Efficient Citer: Tuning Large Language Models for Enhanced Answer Quality and Verification

[1] *Marzieh Tahaei, [1,2]* Aref Jafari, [2] Ahmad Rashid, [1]David Alfonso-Hermelo, [1]Khalil Bibi, [1]Yimeng Wu, [2]Ali Ghodsi, [1]Boxing Chen, [1]Mehdi Rezagholizadeh

[1]Huawei Noah's Ark Lab
[2] David R. Cheriton School of Computer Science, University of Waterloo
{marzieh.tahaei, boxing.chen, mehdi.rezagholizadeh}@huawei.com
{aref.jafari, a9rashid, ali.ghodsi}@uwaterloo.ca

Abstract

In recent years, there has been a growing interest in utilizing external knowledge to reduce hallucinations in large language models (LLMs) and provide them with updated information. Despite this improvement, a major challenge lies in the lack of explicit citations, which hampers the ability to verify the information generated by these models. This paper focuses on providing models with citation capabilities efficiently. By constructing a dataset of citations, we train two model architectures: an FID-style FLAN-T5 model for efficient answer composition and a 13B LLaMA model known for its success in instruction following after tuning. Evaluation on fluency, correctness, and citation quality is conducted through human assessment and the newly introduced Automatic LLMs' Citation Evaluation (ALCE) benchmark. Results demonstrate significant improvements in answer quality and efficiency, surpassing the performance of the popular ChatGPT on some of the metrics. The models exhibit exceptional out-of-domain generalization in both human and automatic evaluation. Notably, the FID-style FLAN-T5 model with only 3B parameters performs impressively compared to the 13B LLaMA model.

The growing popularity of LLMs in information-seeking tasks is undeniable, thanks to their ability to generate fluent, realistic responses. However, there is a growing concern regarding the information accuracy of these responses, and the ability to verify them. Moreover, information is a temporal, ever-changing concept and therefore a model's internal knowledge can quickly become outdated. One possible way to address these concerns, which has gathered a heightened interest, is retrieval-based LLMs, which incorporate external knowledge during both the training and inference stages. However, factual verification of model responses still remains a challenge.

An effective approach to facilitate factual verification involves equipping LLMs with the ability to cite external information. Several commercial systems, including Bing Chat[1], you.com[2], and perplexity[3], have already implemented this approach by leveraging web-based queries to find relevant information, and utilize it to answer specific questions with the relevant citations. However, details of these models are not publicly available.

Some recent works (Nakano et al., 2021; Menick et al., 2022) have attempted to enable LLMs to cite the provided contexts in their response. However, these works fine-tune models which consist of hundreds of billions of parameters (175 billion and 280 billion respectively) and support a range of functions. In contrast, our objective is to develop an efficient answer composition module that can provide informative answers with correct citations, independent of the passage retrieval module.

Recently, Gao et al. (2023) employed in-context learning (ICL), using instructions and demonstrations to facilitate the models' ability to cite relevant context. They applied this approach to various LLMs, including LLaMA (Touvron et al., 2023), Vicuna (instruction-tuned models) (Chiang et al., 2023), and ChatGPT (a closed-source model) (OpenAI, 2022). While ChatGPT is able to provide relatively high-quality answers with relevant citations, for models like LLaMA 13B and Vicuna providing demonstrations alone proved insufficient. Moreover, the use of long demonstrations in ICL increases the prompt length by thousands of tokens, thereby making inference extremely inefficient.

Other studies (Taori et al., 2023; Dettmers et al., 2023; Peng et al., 2023) have demonstrated the effectiveness of fine-tuning LLaMA 13B models using a relatively small training dataset and limited computational resources. This approach has

*Equal Contribution

[1]https://www.bing.com/new?scdexwlcs=1
[2]https://you.com/
[3]https://perplexity.ai/

Findings of the Association for Computational Linguistics: NAACL 2024, pages 4443–4450

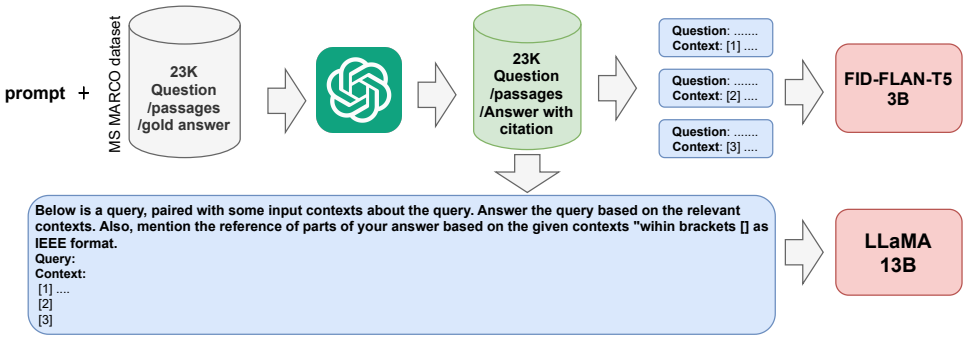

Figure 1: Our data collection and training pipeline for our FIDCiter and LLaMACiter

proven to enhance the models' ability in instruction following and conversation. It can be argued that such capabilities were inherently present in the pretrained models and were unlocked through the fine-tuning process. Building upon this concept, we focus on leveraging a small dataset comprising citations, in the domain of question-answering (QA) to fine-tune LLMs. Specifically, our work makes the following contributions:

- We construct a citation dataset for supervised learning building upon MS MARCO (Nguyen et al., 2016), an open-book QA dataset, by prompting ChatGPT to incorporate citations within the gold responses
- We train an efficient 3 Billion parameter model based on the Flan-T5 (Chung et al., 2022) model with a Fusion-in-Decoder (FID) (Izacard and Grave, 2020), named FIDCiter, and 13B LLaMA model, named LLaMACiter, on this dataset and demonstrate strong performance against ChatGPT and other baselines on both in-domain and out-domain datasets using both human and automatic evaluation.

1 Related work

WebGPT (Nakano et al., 2021), is trained on humans demonstrations on how to use the web environment by issuing commands and answering questions. These demonstrations are used to fine-tune a pretrained GPT-3 (Brown et al., 2020) model. Subsequently, a reward model is employed for reinforcement learning with human feedback (RLHF), which further improves the model's performance.

In contrast, GopherCite (Menick et al., 2022) focuses on reading comprehension by utilizing a large context that includes thousands of tokens from multiple pages. GopherCite also utilizes RLHF to fine-tune the model based on human preferences. Both approaches highlight the significance of larger model sizes, with WebGPT using a 175 billion parameter model and GopherCite employing a 280 billion parameter model.

Attributed question answering (Bohnet et al., 2022) involves answering a question while simultaneously providing a reference to a brief segment of text that supports the answer. Unlike our proposed methods where the link to the reference accompanies the answer directly, in this approach, the model generates the answer along with the reference link as an attribute.

2 Training

2.1 Data construction

The Microsoft Machine Reading Comprehension (MS MARCO) dataset is a comprehensive collection of data designed for machine reading comprehension, question answering, and passage ranking(Nguyen et al., 2016). The dataset consists of about 1M instances of web search queries accompanied by human-generated correct answers, as well as sets of positive and negative passages, sourced from the Bing search engine. The provided answers do not have references to the associated passages, and a significant number of samples contain only one positive passage. we use 18K samples from the MS MARCO dataset, where they have at least two relevant contexts. Therefore, each sample contains a query, relevant and irrelevant contexts , as well as the gold answer without citation. We leverage a large language model to generate answers that cite the factual information from relevant contexts. To achieve this, for each sample, we provide the LLM with a query, a list of relevant contexts (between 2-5), and a gold answer. We then prompt the model to use the gold answer and the provided relevant context to generate a new answer that includes citations in IEEE format. Finally, we add

Table 1: Evaluation of human performance on MS MARCO, MIRACL, ELI5, and ASQA datasets. For MS MARCO and MIRACL, 50 randomly chosen samples are assessed, while 100 randomly selected samples are evaluated for ELI5 and ASQA datasets. Each sample comprises five passages. Criteria evaluated include Fluency (FL), Informativeness (INF), and Citation Quality (CQ).

Model	Prompt Length (words)	MS MARCO			MIRACL			ELI5			ASQA		
		FL	INF	CQ	FL	INF	CQ	FL	INF	CQ	FL	INF	CQ
ChatGPT	1776	**1**	0.76	0.86	**1**	0.56	**0.87**	**0.99**	0.99	**0.88**	**0.97**	0.99	**0.91**
FIDCiter (3B)	10 (2/encoder)	**1**	**0.97**	**0.89**	0.99	0.86	0.76	**0.99**	0.95	0.81	0.86	0.94	0.78
LLaMACiter (13B)	61	0.98	0.95	0.72	0.97	**0.90**	0.79	**0.99**	**1**	0.81	0.89	**1**	0.80

irrelevant contexts to the positive contexts to make the number of contexts equal to 5. This way each query in training data has five contexts from which 2-5 are relevant and the rest are irrelevant. This enforces the model to be able to ignore irrelevant contexts (due to imperfections in IR system) and only use the positive ones for inference. To generate the target answer, we employed Chat GPT, prompting it to generate new answers based on the given passages and the gold answer for each sample while citing different parts of its answer based on the provided passages (see appendix A.3). Subsequently, we removed any hallucinated references and cleaned the citation format used by Chat GPT. The final dataset consists of approximately 18k samples.

2.2 Models

For our 3B model, we adopt an FID-style architecture (Izacard and Grave, 2020) that enables independent passage processing within the encoder while ensuring collective aggregation throughout the decoder. This independent processing allows for the efficient handling of numerous contexts, as it only requires attending to one context at a time. As a result, our model exhibits linear growth in computational requirements rather than a quadratic increase in the encoder's computation. As the backbone encoder-decoder model we use Flan-T5 (Wei et al., 2021) with strong instruction following abilities. We refer to this model as FIDCiter.

Furthermore, in line with the success and popularity of LLaMA for instruction following, we also conduct fine-tuning on a 13B LLaMA model referred to as LLaMACiter. See Appendix A.1 for detailed hyperparameters used during training.

3 Evaluation Criteria

3.1 Human evaluation

Due to the lack of well-studied benchmarks for this task, our main metric in this paper is human evaluation. For the MS Marco dataset (Nguyen et al., 2016), we select 50 held-out samples to serve as our test set. Additionally, to ensure diversity and enable out-of-scope evaluation, we also perform human evaluation on a subset of randomely selected 50 samples from MIRACL dataset (Zhang et al., 2022) and 100 samples from ELI5 (Fan et al., 2019) and ASQA (Stelmakh et al., 2022) datasets. During the evaluation process, the LM receives queries and passages along with a prompt and generates answers while citing relevant passages.

After anonymization, the resulting queries, passages, and generated answers are provided to specialist annotators for evaluation. It is important to note that datasets like MIRACL do not provide gold answers. As a result to ensure reproducibility and fair comparison, we do not provide gold answers to the annotators for any of the samples. Three specialist annotators with over one year of experience in data annotation were hired for the evaluation process. They received task-specific training to ensure consistency and reduce bias. The annotators were compensated at an hourly rate of 25 CAD. During the evaluation, each answer was divided into segments separated by citations, and the metrics were calculated individually for each segment and averaged for each sample. The following evaluation metrics were used to assess our test sets:

- **Informativeness**: The metric evaluates the extent to which a generated answer answers the question. For each segment, if it responds at least partially to the query, we consider it informative and assign a score of 1; otherwise, it is given a score of 0.
- **Fluency**: The metric assesses the naturalness and linguistic correctness of the generated segments. If the segment contains no typographical, morpho-syntactic, or lexical errors, it receives a score of 1, otherwise 0.
- **Supportedness**: This metric measures whether factual claims in the generated seg-

Table 2: Comparative Evaluation on ELI5 and ASQA using ALCE Benchmark with 1000 samples for 3 and 5 passages respectively. Numbers for ChatGPT and Vicuna models are taken from (Gao et al., 2023).

		ELI5				ASQA			
Passages	**Model**	**Prompt Length**	**Correctness**	**Citation Quality**		**Prompt Length**	**Correctness**	**Citation Quality**	
		(words)	(Claim)	Precision	Recall	(words)	(EM)	Precision	Recall
5	ChatGPT	2668	12.0	**51.1**	**50.0**	2550	40.4	**72.5**	**73.6**
	FIDCiter (3B)	10 (2/encoder)	14.3	44.1	44.8	10 (2/encoder)	39.4	62.7	64.2
	LLaMACiter (13B)	61	**16.1**	45.7	34.6	61	**41.8**	56.1	55.2
3	Vicuna (13B)	2668	10.0	15.6	19.6	2550	31.9	51.1	50.1
	FIDCiter (3B)	6 (2/encoder)	15.1	**53.0**	**46.0**	**6** (2/encoder)	40.0	**61.6**	**60.3**
	LLaMACiter (13B)	61	**16.8**	44.6	34.8	61	**41.7**	60.1	60.2

ment can be supported by corresponding quotes. If a cited quote supports the information appearing in the segment it is considered correct. The final score, between 0 and 1, is calculated by dividing the number of correctly cited passages by the total number of passages cited in the segment. If the segment does not cite any passage, it is scored as 0.

4 Results

Table 1 presents the results of our human evaluation on 50 held-out samples from MS MARCO. In the case of ChatGPT, the prompt was designed by the ALCE benchmark (Gao et al., 2023) with four demonstrations. We removed extra information like the title of the URL from the prompt to make it more efficient. The results demonstrate that all models achieve almost ideal fluency. However, when it comes to informativeness, both FIDCiter and LLaMACiter outperform ChatGPT. Regarding supportedness, our FIDCiter model provides the best citation quality, followed by ChatGPT. To assess the generalization of our model to new datasets, we further evaluate its performance on MIRACL, ELI5, and ASQA datasets, utilizing a distinct pipeline for extracting relevant passages. The results clearly indicate that our proposed models outperform ChatGPT in terms of informativeness scores. This is because our informativeness metric is designed to penalize answers that include unnecessary irrelevant contexts. ChatGPT in particular has a tendency to provide such chatty answers (See Appendix A.5). As for citation quality, ChatGPT surpasses both FIDCiter and LLaMACiter. Overall, we observe that the drop in both informativeness and citation quality scores when generalizing to other datasets is less pronounced for LLaMACiter compared to FIDCiter. This demonstrates the higher generalization power of larger models (13B vs 3B). More importantly, our models use very short prompts compared to ChatGPT, making them far more efficient candidates for this task. To further evaluate the model's performance on out-of-domain samples, we utilize correctness, citation precision, and citation recall metrics on 1000 randomly selected samples from the ELI5 and ASQA datasets provided by ALCE benchmark. The results are presented in Table 2. Following the approach of (Gao et al., 2023), we compare the performance of FIDCiter with ChatGPT, LLaMA 13B, and Vicuna 13B models. The prompt used for these baselines is the same as the one used for ChatGPT, which includes four demonstrations from the ELI5 dataset. However, due to the extensive prompt length, Vicuna is limited to only three passages. The results indicate that with three passages. The Vicuna model exhibit very low performance in terms of correctness and citation quality. On the other hand, our FIDCiter model demonstrates significantly better results with a prompt length that is 444 times shorter. Moreover FiDCiter is computationally efficient due to the independent processing of inputs in the encoder. When comparing the performance of our models with ChatGPT using five passages, our LLaMACiter model achieves the highest correctness score among all models.

5 Conclusion

In conclusion, this work improves the reliability and verifiability of question answering by addressing the challenge of explicit citations efficiently. This is achieved by constructing a citation training data and fine-tuning two models: an FID-FLAN-T5 model optimized for answer composition and a 13B LLaMA model. The evaluation, conducted through human assessment and the ALCE benchmark, demonstrates notable improvement in answer quality, and citation accuracy while being far more efficient than the baselines. The presented models surpass the popular ChatGPT and exhibit remark-

able out-of-domain generalization, as evidenced by both human and automatic evaluation. Particularly, the 3B FID-style FLAN-T5 model, despite its smaller size, performs exceptionally well and competes with our 13B LLaMA model. Future research could explore tuning models capable of both question answering and instruction following.

6 Limitations

Our current human evaluation approach can be extended beyond direct assessment to include pairwise comparison, which can help reduce bias and provide more robust evaluations. The automatic evaluation using ALCE also has its own limitations. The accuracy of the NLI model used for assessing citation quality can impact the reliability of the scores obtained. Additionally, the process of generating sub-claims from the gold answer, which is used for computing correctness, can be open-ended and subjective, introducing potential subjectivity into the evaluation process.

In this paper, our training data is merely based on MS MARCO dataset. To further enhance the generalization ability of the proposed models, it would be beneficial to construct training data by combining samples from diverse datasets.

References

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

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

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

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

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages. *arXiv preprint arXiv:2210.09984*.

A Appendix

A.1 Training details

To fine-tune the FID-FLan-T5 model, we set the batch size to 8 and the input(per encoder)/output sequence length to 256/1800 tokens, respectively. We set the learning rate to 5e-5 with a linear decay scheduler and perform warmup during training. We train the model using Adam for 6 epochs, with a total of 5 provided contexts used in the training process. Training on 8 V100 GPUs took approximately 60 hours.

For fine-tune the LLaMA model, we set the batch size to 512 and the sequence length to 1024 tokens. Learning rate was set to 3e-5 with linear decay scheduler and performing warmup similar to FLAN-T5. AdamW optimizer was used for model training and the model was trained for 5 epochs. Five contexts was provided for each sample during the training process and the training was happened on 8 V100 GPUs for approximately 16 hours.

A.2 Inference parameters

To infer the FID-Tlan-T5 model, we utilize the Hugging Face library's generate function. We employ a combination of sampling and beam search, setting the number of beams to 5 and the top P value to 0.95, respectively.

For the LLaMA model, we utilized the Hugging Face library's generate function as well. We used beam search and the number of beams and the top k value is set to 4 and 40 respectively.

A.3 Data collection prompt

For collecting the data samples, we passed the queries and the passages in the MS MARCO dataset as well as the gold answer of each sample and asked the model to generate a new answer by considering the given gold answer and mentioning the passages as the references to the generated parts of the answer. Here you can find the used prompt:

prompt: I will give a question and several context texts about the question as well as the gold answer to the question. By considering the given contexts, the gold answer regenerates an answer to the question. Also, mention the reference of parts of your answer based on the given contexts within brackets [] in IEEE format. Your answer must be similar to the given gold answer.

A.4 Automatic evaluation metric from ALCE Benchmark

There are a few automatic evaluation methods for attributed question answering tasks proposed in the literature (Yue et al., 2023; Gao et al., 2023) In this work, for the automatic evaluation we use correctness, citation recall and precision on ELI5 and ASQA datasets from ALCE (Gao et al., 2023) benchmark. Below is a brief description of each metric.

- **Correctness:** For ELI5 dataset, InstructGPT is used to generate three "sub-claims," and a natural language inference (NLI) model called TRUE is employed to check if the model output entails these sub-claims. This method, known as claim recall, proves to be a more accurate measure of correctness compared to existing metrics.

- **Citation quality:** Citation quality is assessed using two metrics: citation recall and citation precision. Citation recall determines if the output is fully supported by the cited passages, while citation precision identifies any irrelevant citations. ALCE employs the TRUE10 model, fine-tuned on NLI datasets, to automatically determine if the cited passages entail the model generation. The human evaluation demonstrates a strong correlation with the automatic evaluation procedure. For citation recall, each statement's recall is calculated by checking if there is at least one citation and if the concatenated passages entail the statement, using an NLI model.

A.5 Example of our human evaluation

Tables 3, 4, and 5, demonstrate an example of the human evaluation process for a data sample from the Miracle dataset on ChatGPT, FLAN-T5, and

LLaMA models. The sample has a query and five corresponding passages. The evaluators divided the answer of each model into cited sections and evaluated informativeness, citation quality and logic of each section separately. They scrutinized each section within the cited passages to determine the citation quality. The final score was calculated by dividing the number of passages that accurately cited the section by the total number of passages referencing it. You can find the query and the list of the passages of this sample below:

Query: What is the wavelength of visible light? Passages: [1] The visible spectrum is the portion of the electromagnetic spectrum that is visible to the human eye. Electromagnetic radiation in this range of wavelengths is called visible light or simply light. A typical human eye will respond to wavelengths from about 380 to 740 nanometers. In terms of frequency, this corresponds to a band in the vicinity of 430–770 THz.

[2] Electromagnetic radiation with a wavelength between 380 nm and 760 nm (400–790 terahertz) is detected by the human eye and perceived as visible light. Other wavelengths, especially near infrared (longer than 760 nm) and ultraviolet (shorter than 380 nm) are also sometimes referred to as light, especially when the visibility to humans is not relevant. White light is a combination of lights of different wavelengths in the visible spectrum. Passing white light through a prism splits it up into the several colors of light observed in the visible spectrum between 400 nm and 780 nm.

[3] Light, or visible light, is a very narrow range of electromagnetic radiation of a wavelength that is visible to the human eye, or 380–750 nm which equates to a frequency range of 790 to 400 THz respectively. More broadly, physicists use the term "light" to mean electromagnetic radiation of all wavelengths, whether visible or not.

[4] Light is electromagnetic radiation within a certain portion of the electromagnetic spectrum. The word usually refers to visible light, which is the visible spectrum that is visible to the human eye and is responsible for the sense of sight. Visible light is usually defined as having wavelengths in the range of 400–700 nanometres (nm), or 4.00×10 to 7.00×10 m, between the infrared (with longer wavelengths) and the ultraviolet (with shorter wavelengths). This wavelength means a frequency range of roughly 430–750 terahertz (THz). The main source of light on Earth is the Sun. Sunlight provides the energy that green plants use to create sugars mostly in the form of starches, which release energy into the living things that digest them. This process of photosynthesis provides virtually all the energy used by living things. Historically, another important source of light for humans has been fire, from ancient campfires to modern kerosene lamps. With the development of electric lights and power systems, electric lighting has effectively replaced firelight. Some species of animals generate their own light, a process called bioluminescence. For example, fireflies use light to locate mates, and vampire squids use it to hide themselves from prey.

[5] Blue Light is a range of the visible light spectrum, defined as having a wavelength between 400495 nm. This short wavelength means that blue light is a type of high-energy visible light, defined as having a wavelength between 400 and 450 nm. Violet, indigo, and some blue-green light are other types of high-energy visible light.

Segments	Informativeness	Fluency	Citation quality
The wavelength of visible light ranges from about 380 to 740 nanometers, with frequencies in the vicinity of 430-770 THz[1][2][4].	1	1	1
This electromagnetic radiation is the only range that our human eyes can detect[1][3].	0	1	0.5
Specifically, light is defined as electromagnetic radiation within the wavelength range of around 400 to 700 nanometers or frequency range of roughly 430-750 THz[4].	1	1	1

Table 3: ChatGPT output of the sample example. The answer to the given query is broken into 3 pieces

Segments	Informativeness	Fluency	Citation quality
The wavelength of visible light is between 380 nm and 760 nm.[2][4]	1	1	0.5

Table 4: FLAN-T5 output of the sample example. The answer to the given query is broken into 1 section

Segments	Informativeness	Fluency	Citation quality
The wavelength of visible light is between 380 nm and 760 nm or 400–790 THz.[2][1]	1	1	0.5

Table 5: LLaMA output of the sample example. The answer to the given query is broken into 1 section