Verifiable Generation with Subsentence-Level Fine-Grained Citations

Shuyang Cao and **Lu Wang**
University of Michigan
Ann Arbor, MI
{caoshuy, wangluxy}@umich.edu

Abstract

Verifiable generation requires large language models (LLMs) to cite source documents supporting their outputs, thereby improve output transparency and trustworthiness. Yet, previous work mainly targets the generation of sentence-level citations, lacking specificity about which part of the sentence is backed by which cited source. This work studies verifiable generation with subsentence-level fine-grained citations to locate the generated content that is supported by the cited sources in a more precise way. We first present a dataset, SCIFI, comprising 10K Wikipedia paragraphs with subsentence-level citations.[1] Each paragraph in SCIFI is paired with a set of candidate source documents for citation and a query that triggers the generation of the paragraph content. On SCIFI, we then evaluate the performance of state-of-the-a rt LLMs and strategies for processing long documents designed for these models. Our experiment results reveal key factors that can enhance the quality of citations, including the expansion of the source documents' context to be accessible to the models and the implementation of specialized model tuning.

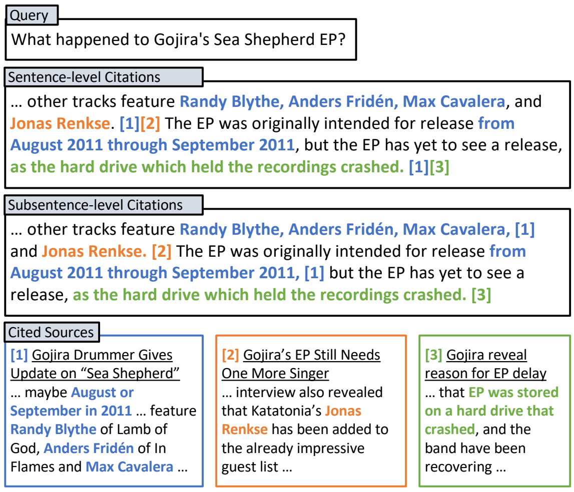

Figure 1: Example of subsentence-level citations. Compared to sentence-level citations, the finer granularity of subsentence-level citations more precisely connect the generated content with the supporting source documents.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in seeking and synthesizing information from given documents (Touvron et al., 2023; OpenAI et al., 2023), and many LLM-powered tools are available to the general public. However, concerns have emerged regarding their outputs' factual accuracy, faithfulness to the source documents, and trustworthiness in general (Zhang et al., 2023; Peskoff and Stewart, 2023).

To address these concerns, recent research has introduced a new generation paradigm, *verifiable generation*, where LLMs are required to include citations to source documents in the model outputs, to support their statements (Bohnet et al., 2023; Gao et al., 2023). Verifiable generation enables users to trace the information back to its source and verify its correctness, enhancing the transparency and reliability of models. Nonetheless, existing work typically provides sentence-level citations. They are unable to indicate which part of the sentence is supported by each referenced source document, leaving the effort to users to infer the connection between the content and its citation (Liu et al., 2023a; Schuster et al., 2023). This ambiguity hinders the user's ability to verify the information efficiently and understand the scope of the supporting documents.

We argue that verifiable generation with a *finer granularity* is critical for further improving the transparency and trustworthiness of LLMs. Compared to sentence-level citations, fine-grained citations, such as subsentence-level annotations, allow for more precise localization of the informa-

[1] Our data is available at https://shuyangcao.github.io/projects/subsentence_citation/.

Findings of the Association for Computational Linguistics ACL 2024, pages 15584–15596
August 11-16, 2024 ©2024 Association for Computational Linguistics

tion that is sourced from the referenced document and support easier assessment of its accuracy, as shown in Figure 1. Furthermore, the importance of fine-grained citations to readers is quantitatively evinced by their prevalence in the popular Wikipedia pages. For instance, pages that are more frequently viewed are more likely to use citations with finer granularity, as demonstrated in Figure 2.

In this work, we aim to investigate LLMs' ability of producing output with fine-grained citations. Specifically, we focus on subsentence-level citations, which are commonly used in information-rich documents like encyclopedias and research papers. To facilitate the study, we first collect **SCIFI**, containing **Sub**sentence-level **Ci**tation of **Fi**ne granularity based on 10K paragraphs from Wikipedia, where rich citation information is available. For each paragraph in SCIFI, documents that are cited in the Wikipedia page where the paragraph belongs to are provided as the citation candidates. We also create a query based on the content of each paragraph to guide LLMs to generate relevant content.

We benchmark state-of-the-art LLMs on our dataset, including OpenAI GPT (OpenAI et al., 2023), Llama2 (Touvron et al., 2023), Vicuna (Zheng et al., 2023), and Mistral (Jiang et al., 2023). To consume the lengthy source documents, we explore three document reading strategies, that respectively target leading context of all source documents, a large portion of the context of all source documents, and full context of selected source documents. For open-source LLMs, we also examine the effect of fine-tuning with the training samples in our dataset. For evaluation, we assess the citation behavior, citation quality, and answer quality of model output. We find that: (1) complete source document context improves citation quality in LLMs; (2) larger model sizes increase answer quality but not citation quality; (3) fine-grained citation generation requires supervised fine-tuning.

Our contributions are summarized as follows: (1) We collect SCIFI which consists of 10K queries paired with reference answers that are rich in subsentence-level fine-grained citations to the source documents, enabling training models for verifiable generation with finer granularity. (2) We analyze performance of state-of-the-art LLMs augmented with various document processing strategies on SCIFI, highlighting directions that could advance model development for verifiable generation.

2 Related Work

Verifiable Generation. Forming in-line citations in verifiable generation challenges LLMs' abilities to ground their generation in source documents and perform accurate attribution. To teach LLMs to include in-line citations in their outputs, early work fine-tunes LLMs with human written demonstrations (Nakano et al., 2022) or model-generated samples verified by human annotators (Menick et al., 2022), but their privately hosted training data prevents follow-up studies. The introduction of more capable LLMs (Jiang et al., 2023; OpenAI et al., 2023) makes it feasible to prompt LLMs with well-crafted instructions to cite source documents during generation (Gao et al., 2023), and such behavior has been activated in online systems that are based on LLMs (Liu et al., 2023a), though the quality of the generated citations leaves large room for improvement (Malaviya et al., 2023). To enhance the citation quality, recent studies consider fetching source documents that better entail the output content (Li et al., 2023) or enabling LLMs to refine its outputs (Sun et al., 2024). Nevertheless, existing work largely focuses on verifiable generation with sentence-level citations, without clearly indicating the exact portion in the output that is supported by the source documents. This work, on the other hand, explores verifiable generation with finer granularity.

Fact-based Evaluation. Evaluation of outputs produced by LLMs is challenging due to their open-ended nature. Recent work resorts to fact-based evaluation, where an output is first decomposed into independent facts and compared against facts in the reference output (Liu et al., 2023b). To circumvent data collection for fact decomposition model training, LLMs have been instructed to extract facts from the output to be evaluated (Kamoi et al., 2023). Furthermore, Min et al. (2023) leverage LLMs' strong capability of identifying content with similar semantics and design an LLM-based fact comparison module for more accurate evaluation. While previous work utilizes fact-based framework for evaluation of factual entailment and precision, we assess both answer coverage and citation quality with fact-based evaluation.

3 SCIFI

To benchmark LLMs on verifiable generation with finer granularity, we first collect SCIFI, a dataset

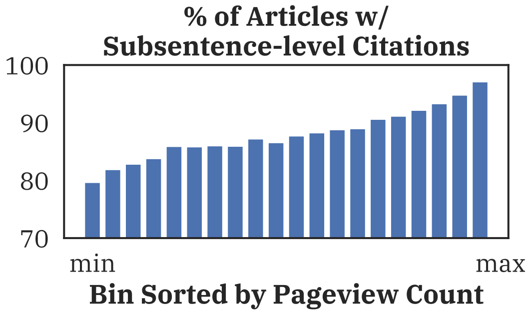

Figure 2: Percentages of Wikipedia articles featuring subsentence-level citations, reported across 20 bins sorted by pageview counts. Popular articles are more likely to include fine-grained citations.

Dataset	# Samples	# Tokens	Density	Fine-grained?
WICE	1,967	27.5	1.88	✗
HAGRID	4,532	40.5	1.23	✗
EXPERTQA	2,177	137.2	1.27	✗
SCIFI	10,000	89.6	1.86	✓

Table 1: Statistics of SCIFI and existing datasets. SCIFI is larger in size, has a high citation density, and supports fine-grained citations.

containing questions that require synthesizing information from multiple source documents to answer and demand more fine-grained citations to precisely locate information supported by different source documents. SCIFI is based on Wikipedia, as Wikipedia articles may densely include citations to support their content.

Before data collection, we examine how prevalent fine-grained citations are used in Wikipedia articles. Out of 100K randomly picked Wikipedia articles, 85% feature subsentence-level citations. We further sort these articles based on their pageview statistics and divide them into 20 bins.[2] For each article, we tallied the total number of views over a five-year period to minimize the impact of transient trends. As can be seen from Figure 2, bins with higher pageview counts show a greater tendency of fine-grained citations, highlighting their importance to readers.

Data Collection. For each article, we extract its text content from the Wikipedia dump while keeping track of the positions and referenced sources of citations in the article. We focus on cited sources that link to downloadable websites and obtain their text content, as it is infeasible to have a uniform process to download and accurately extract content for all types of sources (e.g., images, and PDF files). The position and meta information of all types of cited sources are still preserved, though, to facilitate other research problems such as identifying citation-worthy content.

We further formulate our dataset into a question-answering dataset by creating queries that serve as constraints of the model generated content. Queries are crafted for paragraphs with dense citations rather than entire articles, for more appropriate target lengths to support more sophisticated LLM techniques (e.g., in-context learning). We rank paragraphs with at least 3 citations based on *citation density*, by dividing the total number of citations by the total number of sentences. From the top 25% of these ranked paragraphs, we randomly sample 10K paragraphs to promote paragraphs with more fine-grained citations. For each selected paragraph, we prompt GPT-4 to generate a query asking about the content in the paragraph. Having queries to guide and constrain model outputs allows for more robust evaluation of content quality by anchoring to the reference paragraphs.

Eventually, each sample in SCIFI contains a reference paragraph, the query generated for the paragraph, and all source documents that are cited by the article where the reference paragraph belongs. The dataset is split into training and test sets with 9K and 1K samples.

Statistics. We report statistics of SCIFI along with recent datasets that involve citations in the outputs, including WICE (Kamoi et al., 2023), HAGRID (Kamalloo et al., 2023), and EXPERTQA (Malaviya et al., 2023) in Table 1. As each sample in WICE is a single sentence with citations, for fair comparisons, we compute citation density only for sentences containing at least one citation on all datasets. Our dataset contains samples with high citation density and moderate target lengths. More importantly, it includes citations with finer granularity, while others focus on sentence-level citations. More details of collection and statistics of SCIFI are included in Appendix A.

4 Experiment Setups

Task Setup. Given a sample in SCIFI, the LLM to be tested takes as input the query and candidate source documents to generate a paragraph with

[2]The pageview statistics are obtained via the Wikimedia pageview API: https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews.

fine-grained citations. Due to the sheer volume and excessive length of all available source documents, it is impractical to input them into the LLM simultaneously. Therefore, we provide the LLM with positive source documents—those cited by the reference paragraph, and 5 randomly selected negative source documents that are not cited by the reference paragraph. We shuffle these source documents before feeding them into the LLM.

Models. Though we limit the number of source documents input to the model, their concatenation remains lengthy. To allow the model to process the source documents, we consider (1) truncating source documents to the same size such that their concatenation can be consumed by the model (**Truncated**); and (2) providing summaries of source documents to the model (**Summary**). Additionally, we design a **two-stage** framework that iteratively selects the documents to be covered in next sentence by reading their summaries and writes the next sentence by consuming the original text of the selected documents. This selection process reduces the size of the document pool for the generation phase, thereby allowing the inclusion of more complete document context.

We examine the efficacy of both proprietary and open-source backbone models. As for proprietary models, we test GPT-3.5 with 16K context length and GPT-4 with 8K context length. The 7B and 13B variants of Llama2, 7B variant of Vicuna, and 7B variant of Mistral are chosen for the open-source backbones. We use their RLHF-tuned version for all open-source LLMs. Besides **in-context learning** that is used for all experimented models, we additionally perform **supervised fine-tuning** for open-source models with 4,000 samples in the training set of SCIFI to investigate the effect of fine-tuning.

Evaluation Metrics. We target the assessment of LLMs' ability to produce subsentence-level fine-grained citations, precisely cite the supporting documents, and cover sufficient information for answering the question. We first calculate the **citation density** of model outputs, which is the average number of citations in each output sentence.

For citation quality, we follow Rashkin et al. (2023) and measure how well the output statements entail the cited sources. Unlike sentence-level citations which can be paired with the entire sentence for entailment assessment, evaluation of fine-grained citations requires segmenting the sentence and mapping citations to the sentence portions they support. Inspired by **fact-level entailment** (Min et al., 2023), we decompose model outputs into individual facts and pair citations with facts using heuristics (Appendix C). Following Gao et al. (2023), we run an off-the-shelf entailment model (Honovich et al., 2022) to obtain the entailment levels between generated citations and facts.

Fact-level evaluation is also applied to assess answer quality. For each fact in the reference paragraph, we check if it entails the model output. The aggregation of the fact-level entailment scores reflects the **coverage** of reference facts (i.e., recall).

Strategy	Density	Density (sub)	Citation Ent.	Cover.
GPT-3.5				
Truncated	0.57	0.07	19.57	22.53
Summary	0.59	0.09	29.75	22.08
Two-stage	0.53	0.03	35.30	19.97
GPT-4				
Truncated	0.68	0.20	39.87	**25.86**
Summary	0.84	0.26	47.00	24.53
Two-stage	**1.15**	0.25	**58.56**	21.60
Llama2-7B				
Truncated	0.57	0.17	24.03	18.59
Summary	0.53	0.13	30.06	19.82
Llama2-13B				
Truncated	0.51	0.17	17.87	20.20
Summary	0.49	0.17	21.23	21.91
Vicuna-7B				
Truncated	0.80	0.12	27.48	16.76
Summary	0.64	0.09	30.10	18.69
Mistral-7B				
Truncated	1.06	**0.31**	48.99	20.23
Summary	1.07	**0.31**	49.42	20.20

Table 2: Evaluation results of outputs produced by different strategies and backbone models using in-context learning. (sub): subsentence-level; Ent.: entailment; Cover.: coverage. The best score per metric is in **bold**, while the best strategy per backbone model is underlined. Overall, document reading strategies that provide more complete context yield better citation quality.

5 Results

Source documents with more complete context benefit citation quality, as indicated by the fact-level entailment scores of citations produced by different document reading strategies with in-context learning (Table 2). Summaries can inform models of the major content in the source documents, while truncation only exposes leading content and prevents accurate connection between generated content and supporting documents, thus consis-

Strategy	Density	Density (sub)	Citation Ent.	Cover.
Llama-13B				
Truncated	+0.78	+0.27	+14.22	-2.86
Summary	+0.81	+0.29	+7.94	-5.06
Vicuna-7B				
Truncated	+0.81	+0.60	+1.50	-1.42
Summary	+0.89	+0.54	-0.87	-2.31

Table 3: Improvement of performance after supervised fine-tuning. Negative numbers indicate drops in performance. Supervised fine-tuning encourages model to produce more subsentence-level citations, though not always for citation quality.

tently yielding worse citation quality. Two-stage generation allows for the most complete document context, boosting the citation accuracy, yet its effectiveness relies on strong instruction-following capabilities of the LLMs.[3]

Increasing model sizes promotes answer quality. Across different backbone models and strategies, the coverage of reference facts increases after switching to a larger model within the same family, though larger model sizes do not guarantee an enhancement in citation quality. This reveals that the pre-training designs of different backbone LLMs might all aim for stronger question-answering capabilities, but assign varying significance to their citation and attribution capabilities.

Generation of fine-grained citations requires additional training. We observe that the density of fine-grained citations generated by the same backbone LLMs remains stable across different document reading strategies. By contrast, models generate substantially more fine-grained citations after supervised fine-tuning, as shown in Table 3. However, the effectiveness of supervised fine-tuning in enhancing citation quality varies across models. We think that supervised fine-tuning encourages LLMs' behaviors of generating fine-grained citations. Yet, the development of LLMs' abilities to correctly link sentence parts with supporting documents requires more specialized and sophisticated training procedure, which highlights the challenge of this task. Future directions may include design builtin citation or attribution mechanisms during LLM pretraining (Khalifa et al., 2024).

[3]The two-stage strategy is only adopted by the OpenAI GPT families, as other models could not consistently follow the output format designed for the strategy, resulting in invalid results.

6 Conclusions

We study verifiable generation with subsentence-level fine-grained citations. SCIFI, a benchmark containing 10K subsentence-level citation-rich paragraphs together with candidate cited sources and queries, is collected to support the training and evaluation of models on this task. On SCIFI, experiments with state-of-the-art LLMs and various processing strategies demonstrate the importance of source document context and training with citation-rich data.

Acknowledgments

This work is supported in part by the National Science Foundation through grant IIS-2046016. Shuyang Cao is supported by a Bloomberg Data Science Ph.D. Fellowship. We thank ARR reviewers for their feedback.

Limitations

In our experiments, we employ three strategies for LLMs to handle lengthy source documents and observe improved performance when the strategy enables a more comprehensive context. Yet, more sophisticated strategies can be designed, and techniques that expand input limits of LLMs can be explored, which could potentially lead to higher performance in our benchmark study.

When evaluating the attribution quality, we leverage an existing entailment model tailored for assessing the entailment relation between short passages. However, we use it to measure the entailment relation between an extracted fact and a long source document. Although we follow the technique in previous work (see Appendix C.3) to extend the application of the off-the-shelf entailment models to long documents, more accurate evaluation can be achieved by developing entailment models specialized for long documents.

Ethical Considerations

Our benchmark enables the evaluation of LLMs' ability to generate subsentence-level citations. With subsentence-level citations, LLM developers are able to present LLM outputs that contain precise pointers directing users to supporting sources of different parts of output sentences. While this would enhance user trust in LLMs, it is worth noting that our dataset comprises texts that are formally written in Wikipedia and the candidate supporting documents are from reliable online sources.

An LLM with outstanding performance on our dataset might cite documents with fake facts if the candidate documents are from unreliable sources, which might further propagate incorrect information. Users of our benchmark should also consider the reliability of their candidate supporting documents when examining the reliability of LLM-based applications.

References

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

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

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. 2024. Source-aware training enables knowledge attribution in language models. *arXiv preprint arXiv:2404.01019*.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2023. Llatrieval: Llm-verified retrieval for verifiable generation.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers.

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Denis Peskoff and Brandon Stewart. 2023. Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438, Toronto, Canada. Association for Computational Linguistics.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, 49(4):777–840.

Tal Schuster, Adam D. Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William W. Cohen, and Donald Metzler. 2023. Semqa: Semi-extractive multi-source question answering.

Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2024. Towards verifiable text generation with evolving memory and self-reflection.

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

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei

Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

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

A Details of Data Collection

A.1 Wikipedia Article and Cited Source Collection

Wikipedia paragraphs in SCIFI are extracted from the Aug 20, 2021 Wikipedia dump. For each source that links to a downloadable website, we retrieve its HTML file from Internet Archive[4] and extract its metadata and text content using Trafilatura.[5]

A sample example is shown in Table 13.

Our dataset will be made publicly available under the CC BY 4.0 license.[6]

A.2 Topic Distribution

We use the topic model provided by Wikimedia[7] to determine the topic of the page from which each paragraph in SCIFI is extracted. For each page, we select its top 3 predicted topics, considering that a single page could cover multiple topics. Top 10 topics are presented in the Table 4. Paragraphs in our dataset come from pages of diverse topics.

Topic	Percentage
STEM.STEM	39.45
Culture.Media.Media	23.42
Geography.Regions.Europe	17.77
Culture.Biography	16.97
STEM.Technology	12.95
Geography.North America	12.54
Geography.Asia	9.13
History and Society.Politics and Government	7.78
Culture.Literature	6.01
Culture.Philosophy and Religion	5.67

Table 4: Top 10 topics covered by the pages where samples in SCIFI are extracted.

A.3 Additional Statistics

SCIFI has 39292 sentences in total, 30.1% of which have more than one citation. As subsentence-level citations frequently occur at the end of clauses marked with punctuation, we also check the percentage of subsentence-level citations in SCIFI that are not attached to punctuation. We find that 30.2% of subsentence-level citations are not located around punctuation, indicating a decent level of diversity in subsentence-level citations.

A.4 Prompt for Query Generation

We use GPT-4 to create query for each paragraph in SCIFI. The prompt we use is shown in Table 5.

B Additional Results

We additionally test the model performance in the oracle setup, where only positive source documents (i.e., those cited by the reference paragraph) are fed into the LLM (Table 6). In the oracle setup, trends of the results are similar to those in the regular setup, with all models tending to produce answers of high quality while maintaining the citation quality. This indicates that removing source documents irrelevant to the query offers minimal help to LLMs for verifiable generation.

C Details of Experiment Setups

C.1 Model Prompting

We use 2-shot examples for all experiments with in-context learning. Prompts for the **Truncated** and **Summary** strategy are shown in Table 7 and 9. Summaries of the source documents are generated by GPT-3.5 with 16K context length using the prompt in Table 8.

In the **Two-stage** strategy, the LLM is given two different prompts to perform document selection and answer sentence generation. When selecting source documents, the LLM is informed of the current answer and all its previous selections, as shown in Table 10. When generating the next answer sentence, the LLM is provided the current answer and the original text of the selected source documents (Table 11).

All models are only prompted once due to the API cost.

C.2 Supervised Fine-tuning

We use `LLaMA-Factory`[8] for fine-tuning Llama2 models in our experiments. We adopt LoRA (Hu et al., 2021) and fine-tune the model for 3 epochs with a learning rate of 5e-5 and a batch size of 16. All LoRA-compatible projection layers are tuned,

[4] http://web.archive.org/
[5] https://github.com/adbar/trafilatura
[6] https://creativecommons.org/licenses/by/4.0/
[7] https://meta.wikimedia.org/wiki/Machine_learning_models/Production/English_Wikipedia_article_topic
[8] https://github.com/hiyouga/LLaMA-Factory/

```
Read the following paragraph from a Wikipedia page and create a question whose
answer covers most of the information in the paragraph. The section title and page
title where the paragraph is found are also provided (if any). Try not to create
a compound question.
Page Title: {page}
Section Title: {section}
Paragraph: {paragraph}
Question:
```

Table 5: Prompt for query generation.

Strategy	Density	Density (sub)	Citation Ent.	Cover.
GPT-3.5				
Truncated	0.49	0.06	22.68	22.53
Summary	0.52	0.05	25.60	24.25
Two-stage	0.93	0.05	55.14	19.43
GPT-4				
Truncated	0.66	0.19	38.89	**29.85**
Summary	0.86	0.25	47.25	24.44
Two-stage	**1.37**	**0.28**	**64.11**	20.91
Llama2-7B				
Truncated	0.49	0.10	26.71	23.76
Summary	0.50	0.10	31.16	23.42
Llama2-13B				
Truncated	0.46	0.13	21.48	25.96
Summary	0.48	0.15	24.01	25.24

Table 6: Evaluation results of outputs produced by different strategies and backbone models using in-context learning in the oracle setup. (sub): subsentence-level; Ent.: entailment; Cover.: coverage. The best score of each metric among **bolded.**, while the best strategy for each backbone model is underlined.

with a rank of 32 and a α of 64. Training of each model is conducted using 2 Nvidia A40 GPUs and takes 4 hours to complete.

C.3 Evaluation Metrics

We leverage fact-level entailment to evaluate the citation quality and answer quality. We follow the prompts in previous work (Min et al., 2023; Kamoi et al., 2023) and use GPT-3.5 to conduct fact decomposition for each output sentence separately.

To map citations to an extracted fact, we first map the extracted fact back to a segment in the original sentence. For accurate mapping, we again use GPT-3.5 to identify segments in the original sentence that best represents the extracted fact, with the prompt in Table 12. We then rank generated citations based on their distances to the sentence segment associated with the extracted fact. If two citations have the same distance to the end of the sentence segment, the one after the sentence segment is ranked higher, as we hypothesize that the citation supporting a fact is likely to occur after its completion in the sentence. The top-ranking citation is mapped to the extracted fact.

When evaluating the entailment level, directly pairing the cited source document with the extracted fact is infeasible due to the length of the cited document. Following Kamoi et al. (2023), we divide the cited document into chunks of 256 tokens, calculate the extracted fact's entailment level against each chunk, and take highest entailment level as the final entailment score.

```
Instruction: Write an accurate, engaging, and concise answer for the given question
using only the provided search results (some of which might be irrelevant) and
cite them properly. Use an unbiased and journalistic tone. Always cite after the
completion of each individual fact in the answer. Facts might be completed in the
middle of a sentence.

Question: {query}

Document [1] (Title: {document1_title})
{truncated_document1_text}

...

Document [N] (Title: {documentN_title})
{truncated_documentN_text}

Answer:
```

Table 7: Prompt for generation with the **Truncated** strategy.

```
Summarize the following document within 100 words. Try to keep all the important
dates, numbers, and names.
Title: {title}
Text: {text}
Summary:
```

Table 8: Prompt for article summary generation.

```
Instruction: Write an accurate, engaging, and concise answer for the given question
using only the provided search results (some of which might be irrelevant) and cite
them properly. You are provided summaries of the search results, rather than the
original search results. Use an unbiased and journalistic tone. Always cite after
the completion of each individual fact in the answer. Facts might be completed in
the middle of a sentence.

Question: {query}

Document [1] (Title: {document1_title})
{summary_document1_text}

...

Document [N] (Title: {documentN_title})
{summary_documentN_text}

Answer:
```

Table 9: Prompt for generation with the **Summary** strategy.

```
Instruction: Write an accurate, engaging, and concise answer for the given question
using only the provided search results (some of which might be irrelevant) and cite
them properly. You are provided summaries of the search results, rather than the
original search results. Answer the question sentence by sentence. Now, given the
empty or already written answer, choose which document(s) in the search results
to use for the next sentence of the answer. You can also decide to stop the answer
with [STOP] if you think the answer is complete.

Question: {query}

Document [1] (Title: {document1_title})
{summary_document1_text}

…

Document [N] (Title: {documentN_title})
{summary_documentN_text}

Written Answer Sentences: {current_answer_iter1}
Chosen Documents: {chosen_document_iter1}

…

Written Answer Sentences: {current_answer_iterN}
Chosen Documents:
```

Table 10: Prompt for the selection phase of the **Two-stage** strategy.

```
Instruction: Write an accurate, engaging, and concise answer for the given question
using only the provided search results (some of which might be irrelevant) and cite
them properly. Answer the question sentence by sentence, and the already written
answer sentences are given. Now, write the next sentence of the answer. Use an
unbiased and journalistic tone. Always cite after the completion of each individual
fact in the answer. Facts might be completed in the middle of a sentence.

Question: {query}

Document [{selected_document1_index}] (Title: {selected_document1_title})
{selected_document1_text}

…

Document [{selected_documentN_index}] (Title: {selected_documentN_title})
{selected_documentN_text}

Written Answer Sentences: {current_answer}
Next Sentence:
```

Table 11: Prompt for the generation phase of the **Two-stage** strategy.

```
Find the shortest possible segment in the sentence that reflects the claim. The
segment must be a contiguous substring of the sentence.
Sentence: He made his acting debut in the film The Moon is the Sun's Dream (1992),
and continued to appear in small and supporting roles throughout the 1990s.
Claim: He made his acting debut in 1992.
Segment: 1992

Find the shortest possible segment in the sentence that reflects the claim. The
segment must be a contiguous substring of the sentence.
Sentence: In 1963, Collins became one of the third group of astronauts selected by
NASA and he served as the back-up Command Module Pilot for the Gemini 7 mission.
Claim: Collins became one of the third group of astronauts selected by NASA in
1963.
Segment: In 1963, Collins became one of the third group of astronauts selected by
NASA

Find the shortest possible segment in the sentence that reflects the claim. The
segment must be a contiguous substring of the sentence.
Sentence: A previous six time winner of the Nations' Cup, Sebastian Vettel became
Champion of Champions for the first time, defeating Tom Kristensen, who made the
final for the fourth time, 2-0.
Claim: Sebastian Vettel is a previous six-time winner of the Nations' Cup.
Segment: A previous six time winner of the Nations' Cup

Find the shortest possible segment in the sentence that reflects the claim. The
segment must be a contiguous substring of the sentence.
Sentence: A previous six time winner of the Nations' Cup, Sebastian Vettel became
Champion of Champions for the first time, defeating Tom Kristensen, who made the
final for the fourth time, 2-0.
Claim: Tom Kristensen made the final for the fourth time.
Segment: Tom Kristensen, who made the final for the fourth time

Find the shortest possible segment in the sentence that reflects the claim. The
segment must be a contiguous substring of the sentence.
Sentence: {sentence}
Claim: {fact}
Segment:
```

Table 12: Prompt for fact mapping.

Question: What were the circumstances and details of Richard Blumenthal's military service during the Vietnam War?

Reference: Blumenthal received five draft deferments during the Vietnam War, [5] first educational deferments, then deferments based on his occupation. [1] With part-time service in the reserves or National Guard generally regarded as an alternative for those wishing to avoid service in Vietnam, in April 1970 Blumenthal enlisted in the United States Marine Corps Reserve. He served in units in Washington, D.C., and Connecticut from 1970 to 1976, [2] attaining the rank of sergeant. [4]

Document [1] (Title: Dick Blumenthal, Reporting for Duty)
Perhaps John Kerry, the former junior senator from Massachusetts, did not serve as heroically in Vietnam as he would like us to think. Certainly he wasn't there for very long. But at least he put in an appearance. The same can't be said of Sgt. Richard Blumenthal, a fellow Democrat and the attorney general of Connecticut, who is seeking to become that state's junior senator ...

Document [2] (Title: Blumenthal an easy victor)
HARTFORD – Democratic delegates overwhelmingly and unsurprisingly nominated Attorney General Richard Blumenthal on Friday night to run for the U.S. Senate seat that will be vacated by U.S. Sen. Chris Dodd ...

Document [3] (Title: David Blumenthal, M.D., M.P.P.)
David Blumenthal, M.D., M.P.P., is president of The Commonwealth Fund, a national philanthropy engaged in independent research on health and social policy issues ...

Document [4] (Title: Senator Blumenthal honored at Yale Graduate School diversity conference)
The ninth annual Bouchet Leadership Conference on Diversity in Graduate Education took place at Yale March 30-31. The focus of this year's conference was "Determining the Future of Diversity Discussions." U.S. Senator Richard Blumenthal '73 J.D. (D-CT), who received this year's Bouchet Leadership Award, at the conference, delivered the keynote address ...

Document [5] (Title: Senate hopeful Richard Blumenthal addresses report he lied about Vietnam record)
Connecticut Attorney General Richard Blumenthal (D) was alternately apologetic and defiant Tuesday as he battled to deflect a potentially devastating blow to his Senate campaign: an accusation that he had exaggerated his military service record ...

Document [...] (Title: ...)
...

Table 13: Sample of SciFi. First paragraphs of the first 5 documents in the candidate document pool are shown.