Generation-Time vs. Post-hoc Citation: A Holistic Evaluation of LLM Attribution

Anonymous Author(s)
Affiliation
Address
email

Abstract

Trustworthy Large Language Models (LLMs) must cite human-verifiable sources in high-stakes domains such as healthcare, law, scientific research, where even small errors can have severe consequences. Practitioners and researchers face a choice: let models generate citations during decoding, or let models draft answers first and then attach appropriate citations. To clarify this choice, we introduce two paradigms: *Generation-Time Citation* (`G-Cite`), which produces the answer and citations in one pass, and *Post-hoc Citation* (`P-Cite`), which adds or verifies citations after drafting. We conduct a comprehensive evaluation from zero-shot to advanced retrieval-augmented methods across four popular attribution datasets, and provide evidence-based recommendations that weigh trade-offs across use cases. Our results show a consistent trade-off between coverage and citation correctness, with retrieval as the main driver of attribution quality in both paradigms. `P-Cite` methods achieve high coverage with competitive correctness and moderate latency, whereas `G-Cite` methods prioritize precision at the cost of coverage and speed. We recommend a retrieval-centric, `P-Cite`-first approach for high-stakes applications, reserving `G-Cite` for precision-critical settings such as strict claim verification. Our codes and human evaluation results are available at `https://anonymous.4open.science/r/Citation_Paradigms-BBB5/`

1 Introduction

Just as humans cite sources to demonstrate credibility in their communication, LLM-based AI systems must provide attribution to build trust in their outputs [Phukan et al., 2024, Li et al., 2023]. Trustworthy AI has become a national priority following the White House's Executive Order on AI [House, 2025], and this will become increasingly critical as models scale. Researchers and practitioners working in high-stakes applications need confidence that their LLM-based AI tools are reliable and transparent [Leyli-abadi et al., 2025, Kowald et al., 2024]. To contextualize, consider the legal domain where summarizing lengthy documents is routine. In such settings, it is essential to attribute each generated sentence to its corresponding source text to ensure reliability and transparency [Batista et al., 2025a]. Viewing attribution as a practical pathway to trustworthiness, we categorize existing research into two fundamental paradigms `G-Cite` and `P-Cite`, providing researchers and practitioners with actionable choices for developing attribution-capable LLMs.

Generation-Time Citation (`G-Cite`) creates the text and citation markers together in a single step. *Post-hoc Citation* (`P-Cite`) works differently; it first creates a draft, then adds or checks citations in a separate step. The key difference between these approaches is timing of citations. Beside timings, these approaches differ in how they work technically. `G-Cite`

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

makes citation choices during the normal left-to-right text generation process. It decides *locally* based on what has been written so far and any retrieved evidence. In contrast, `P-Cite` separates citation from text generation entirely. It runs a second pass to examines the complete draft and available evidence to add or verify citations throughout the entire text.

As of now there is no principle and systematic study to compare these two paradigm to determine its limitation and capabilities. Each method is designed differently and evaluated using a different evaluation metric on different datasets. These gap makes it difficult for researchers and practitioners to choose the right method or design better attribution systems. To address this problem, our paper provides a rigorous empirical comparison of both paradigm using common datasets and evaluation metric across diverse categories of methods.

Dataset	Citation granularity	Nativeness	Instances (#)
ALCE	Doc, Sent	`G-Cite`	3,000
LongCite	Sent	`G-Cite`	1,000
REASONS	Doc, Sent	`P-Cite`	12,723
FEVER	Sent, Claim	`P-Cite`	185,445

Table 1: **Details of Datasets used in the Experiments.** FEVER focus on factual verification of claims in health, law and other domains. REASONS is focused on scientific research, and LongCite/ALCE are open domain QA with long and short context respectively. “Doc”, “Sent‘”, “Claim” denotes if the citation is at document-level, sentence-level, or Claim-level.

We benchmark the state-of-the-art methods from both paradigms. Because several existing datasets are tailored to a single paradigm, we adapt them to enable fair, cross-paradigm comparison. More details on the datasets are available in Table 1. We consider four types of methods: Zero-shot, Fine-tuned, Retrieval Augmented Generation (RAG) and two recent approaches from `G-Cite` and `P-Cite` as Advanced methods. Here zero-shot and RAG act as common baselines. For Advanced and Fine-tuned methods, within `G-Cite`, we use CoT Citation [Ji et al., 2024] and LongCite (8b) [Zhang et al., 2025] and within `P-Cite`, we evaluate CiteBART [Çelik and Tekir, 2025] and CEG [Li et al., 2024a]. We assess performance using established quantitative attribution metrics: Citation Correctness, Precision, Recall, Coverage, and Latency. These metrics are widely used in prior work and enable consistent comparison across methods. We also conduct a human evaluation with n=100 instances per method-dataset pair, providing 80% statistical power to detect medium effect sizes (Cohen's d ≥ 0.5) at $\alpha = 0.05$ significance level, for more details, see section 2.

Findings: By analyzing results shown in Figure 1 and Figure 3, we establishes four primary insights: (1) Retrieval augmentation is fundamental as it provides the largest gains in both citation correctness and coverage regardless of paradigm choice; (2) `P-Cite` methods achieve higher coverage with competitive citation correctness as they include more ground-truth citations while maintaining reasonable citation precision and citation recall balance; (3) Advanced methods enable targeted optimization as they allow practitioners to adjust the citation precision and coverage with latency costs; and (4) Organizations should view fine-tuning as an optimization enhancement rather than a replacement strategy, while domain-specific models can improve efficiency and task alignment, retrieval-augmented approaches remain essential to maintain the citation correctness and coverage standards when information accuracy is non-negotiable.

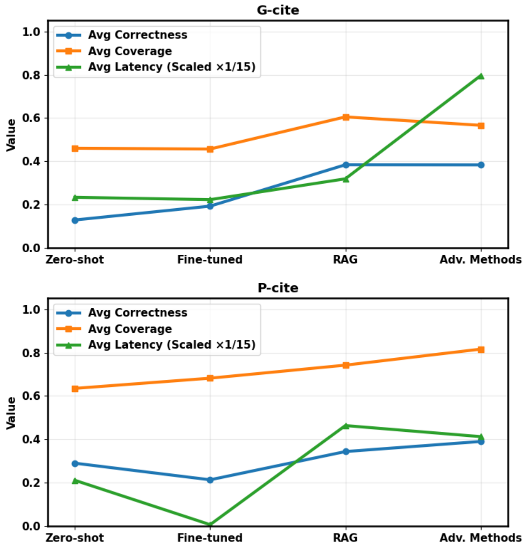

Figure 1: **Citation Quality Trends.** Average citation correctness, entailed coverage, and latency across categories (Zero-shot, Fine-tuned, RAG, Advanced) for the `G-Cite` and `P-Cite` paradigms, averaged over all datasets.

2 Experimentation

We benchmark state-of-the-art methods from both paradigms, `G-Cite` and `P-Cite`, eight in total (four per paradigm). For a fair evaluation, we use *LLaMa-3.1-8B-Instruct* for all methods except for CiteBART which makes use of the BART model.

Datasets. Table 1 shows the datasets that span open-domain QA, scientific citation, and fact verification; long and short context; `G-Cite`-native and `P-Cite`-native settings; and sentence- and document-level granularity. For fair comparison, we adapt each dataset to also support the non-native paradigm. For ALCE [Gao et al., 2023] and LongBench-Cite (both `G-Cite`-native, where the model outputs an answer with inline markers), we create a citation-free draft answer and then let `P-Cite` attach inline citations. For REASONS (`P-Cite`-native), we evaluate `G-Cite` by providing a constrained candidate pool (titles) and prompting the model to rewrite the target sentence with an inline citation. For FEVER [Thorne et al., 2018], we treat each claim as a "needs citation" unit: `G-Cite` generates the statement with an inline citation, while `P-Cite` attaches a citation to the fixed statement.

Methods. We evaluate `G-Cite` and `P-Cite` using a systematic categorization of four method types: (i) Zero-shot prompting, (ii) Fine-tuned models, (iii) Retrieval-augmented methods, and (iv) Advanced hybrid techniques. Zero-shot prompting was used across both the `G-Cite` and `P-Cite` paragraph, serving as a fundamental baselines without task-specific training. Implementation details and zero-shot prompts with examples are in subsection A.2. *Fine-tuned models* represent specialized architectures trained on domain-specific datasets such as LongCite-8B for citation generation and CiteBART for paper citation, both are trained using supervised learning on ALCE and arXiv datasets. Retrieval-augmented methods combine retrieval systems with generative models to enhance citation accuracy and coverage. We considered retrieval-based methods in both `G-Cite` and `P-Cite`. Advanced hybrid techniques integrate multiple methodological components: CoT Citation combines evidence retrieval with chain-of-thought prompting and includes an evidence-insurance step for comprehensive citation coverage, while Citation Evidence Generation (CEG) employs an iterative approach that retrieves relevant information before systematically attaching and verifying citations.

Metrics and Human Evaluation. We use standard citation metrics for evaluation. Specifically, we use five standard metrics: (i) Citation Precision to measure the fraction of correct citations among those produced; ii) Citation Recall to measure the fraction of ground-truth citations that are retrieved;(iii) Citation Correctness to measure the harmonic mean of precision and recall [Aly et al., 2024]; (iv) Coverage to measure the proportion of ground-truth citations present in the generated response [Aly et al., 2024]; (v) Latency to measure the average time (in seconds) taken by each method per dataset instance. We conducted a human assessment using two expert annotators from the university library with experience in AI-assisted citation verification. We evaluated 100 instances per method-dataset pair across two critical quality measures for each generated response, achieving κ = 0.873 inter-annotator agreement. *Answer Correctness* is a strict metric which allow human evaluators to verify whether the provided evidence actually supports each claim made in the generated text. We assign a score of 1 when all claims are properly backed by their cited sources, and 0 when any claims lack adequate support. *Citation Hallucination* allows human evaluators to check whether each citation corresponds to a real source from the reference dataset. Humans assign a score of 1 when citations are fabricated or point to sources outside the ground truth collection, and 0 when all citations are legitimate and verifiable.

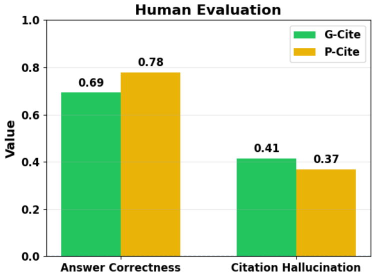

Figure 2: **Human Evaluation Results.** We report *Answer Correctness* (↑), and *Citation Hallucination* (↓), values are averaged over all datasets and methods within each paradigm (`G-Cite` and `P-Cite`).`P-Cite` based methods tends to provide more correct answers with lesser hallucination.

3 Results and Analysis

Our evaluation reveals distinct performance characteristics between the two citation paradigms across all datasets. `P-Cite` consistently achieve higher coverage while maintaining competitive citation correctness compared to `G-Cite`. For practitioners deploying LLMs in information-seeking applications, where users need comprehensive source attribution to verify claims across multiple documents, `P-Cite` methods provide a critical advantage by ensuring broader citation coverage without compromising accuracy. Further, as shown in Figure 2, the human evaluation reinforces these findings: averaged over datasets and

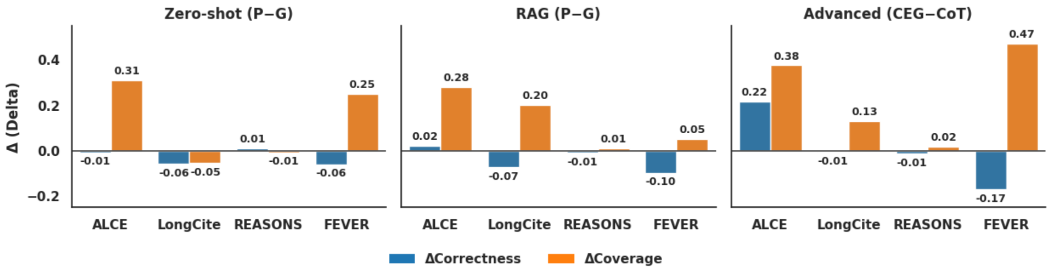

Figure 3: **Coverage and correctness deltas (P-Cite - G-Cite) across datasets.** Positive values indicate P-Cite outperforms G-Cite on the metric; negative values indicate otherwise.

methods, P-Cite shows higher answer correctness than G-Cite (78% vs. 69%) and lower citation hallucination (37% vs. 41%). See Appendix A.3 for detailed results.

Finding 1: On ALCE, the advanced P-Cite achieve 75% coverage with 42% correctness, substantially outperforming the advanced G-Cite which reaches 37% coverage and 21% correctness. Similarly, on LongBench-Cite, P-Cite attains 78% coverage with 12% correctness, while G-Cite achieves 65% coverage with 12% correctness. This shows that in complex information synthesis tasks, such as research summaries, technical reports, or knowledge synthesis, practitioners can expect P-Cite methods to provide citations for approximately twice as many relevant sources as G-Cite methods, dramatically improving the verifiability and trustworthiness of generated content.

Finding 2: On the scientific REASONS dataset both paradigms achieve comparable correctness (P-Cite 26% vs. G-Cite 27%) and near-ceiling coverage (P-Cite 99% vs. G-Cite: 97%). On the FEVER dataset, G-Cite achieves the highest precision and correctness (94%) but limited citation coverage (27%), while P-Cite provides a more balanced profile with high coverage (74%) and strong correctness (75%). This shows that in scientific literature tasks, both P-cite and G-cite methods perform similarly well. However, for fact verification tasks, which are important for legal and policy-driven research, P-cite methods excel when you need comprehensive evidence from multiple sources, while G-cite methods work better when you need extremely precise validation of individual claims.

Finding 3: Retrieval augmentation emerges as the primary driver of citation accuracy. The transition from zero-shot to RAG yields the most substantial and consistent improvements across both paradigms and all datasets. On FEVER, G-Cite correctness improves by approximately 50 percentage points (from 27% to 77%), while on LongBench-Cite, coverage increases by approximately 47 percentage points (from 11% to 58% for G-Cite). Our results clearly suggest that organizations deploying LLMs for information-critical applications should prioritize investment in retrieval infrastructure as the foundational requirement to gain access to relevant, high-quality source material for LLM-based applications.

Finding 4: Advanced methods built on retrieval foundations adjust the citation coverage and latency trade-off. P-Cite delivers high coverage and correctness with moderate latency costs, whereas G-Cite delivers better performance but substantially increases latency. Practitioners must balance operational efficiency with accuracy, where P-Cite offer a practical solution and G-Cite could be used for verification purposes. Fine-tuned models provide incremental improvements but cannot replace retrieval for maintaining content accuracy. Human evaluators, representing end users who rely on generated content for critical decisions, consistently rate P-Cite outputs as more accurate and trustworthy.

4 Conclusion

In this work, we empirically evaluated G-Cite and P-Cite, highlighting their respective capabilities and limitations. Our findings show that retrieval-based attribution is fundamental regardless of paradigm. Using common metrics and datasets, we demonstrate that P-Cite is better suited for high-stakes applications due to higher factuality, while G-Cite is preferable in precision-critical settings. To facilitate future research, we release our code and human evaluations for reproducibility.

References

Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. Peering into the mind of language models: An approach for attribution in contextual question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11481–11495, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.682. URL https://aclanthology.org/2024.findings-acl.682/.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Comput. Surv.*, 55(9), January 2023. ISSN 0360-0300. doi: 10.1145/3555803. URL https://doi.org/10.1145/3555803.

White House. Preventing Woke AI in the Federal Government — whitehouse.gov. https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/, 2025. [Accessed 03-09-2025].

Milad Leyli-abadi, Ricardo J. Bessa, Jan Viebahn, Daniel Boos, Clark Borst, Alberto Castagna, Ricardo Chavarriaga, Mohamed Hassouna, Bruno Lemetayer, Giulia Leto, Antoine Marot, Maroua Meddeb, Manuel Meyer, Viola Schiaffonati, Manuel Schneider, and Toni Waefler. A conceptual framework for ai-based decision systems in critical infrastructures, 2025. URL https://arxiv.org/abs/2504.16133.

Dominik Kowald, Sebastian Scher, Viktoria Pammer-Schindler, Peter Müllner, Kerstin Waxnegger, Lea Demelius, Angela Fessl, Maximilian Toller, Inti Gabriel Mendoza Estrada, Ilija Šimić, et al. Establishing and evaluating trustworthy ai: overview and research challenges. *Frontiers in Big Data*, 7:1467222, 2024.

João Eduardo Batista, Emil Vatai, and Mohamed Wahib. Think before you attribute: Improving the performance of llms attribution systems. *arXiv preprint arXiv:2505.12621*, 2025a.

Bin Ji, Huijun Liu, Mingzhe Du, and See-Kiong Ng. Chain-of-thought improves text generation with citations in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18345–18353, Mar. 2024. doi: 10.1609/aaai.v38i16.29794. URL https://ojs.aaai.org/index.php/AAAI/article/view/29794.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5098–5122, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.264. URL https://aclanthology.org/2025.findings-acl.264/.

Ege Yiğit Çelik and Selma Tekir. Citebart: Learning to generate citations for local citation recommendation, 2025. URL https://arxiv.org/abs/2412.17534.

Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. Citation-enhanced generation for LLM-based chatbots. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1451–1466, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.79. URL https://aclanthology.org/2024.acl-long.79/.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Rami Aly, Zhiqiang Tang, Samson Tan, and George Karypis. Learning to generate answers with citations via factual consistency models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11876–11896, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.641. URL https://aclanthology.org/2024.acl-long.641/.

Martin Funkquist, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. CiteBench: A benchmark for scientific citation text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7337–7353, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.455. URL https://aclanthology.org/2023.emnlp-main.455/.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. Learning to plan and generate text with citations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11397–11417, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.615. URL https://aclanthology.org/2024.acl-long.615/.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. Training language models to generate text with citations via fine-grained rewards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2926–2949, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.161. URL https://aclanthology.org/2024.acl-long.161/.

Yung-Sung Chuang, Benjamin Cohen-Wang, Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James R. Glass, Shang-Wen Li, and Wen tau Yih. Selfcite: Self-supervised alignment for context attribution in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=rKi8eyJBoB.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. Ground every sentence: Improving retrieval-augmented LLMs with interleaved reference-claim generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 969–988, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.55. URL https://aclanthology.org/2025.findings-naacl.55/.

Xiangci Li, Yi-Hui Lee, and Jessica Ouyang. Cited text spans for scientific citation text generation. In Tirthankar Ghosal, Amanpreet Singh, Anita Waard, Philipp Mayr, Aakanksha Naik, Orion Weller, Yoonjoo Lee, Shannon Shen, and Yanxia Qin, editors, *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 90–104, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.sdp-1.9/.

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

Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.566. URL https://aclanthology.org/2022.emnlp-main.566/.

Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. QAMPARI: A benchmark for open-domain questions with many answers. In Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz, editors, *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–110, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.gem-1.9/.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL https://aclanthology.org/P19-1346/.

Yash Saxena, Deepa Tilwani, Ali Mohammadi, Edward Raff, Amit Sheth, Srinivasan Parthasarathy, and Manas Gaur. Attribution in scientific literature: New benchmark and methods, 2025. URL https://arxiv.org/abs/2405.02228.

Harsh Maheshwari, Srikanth Tenneti, and Alwarappan Nakkiran. CiteFix: Enhancing RAG accuracy through post-processing citation correction. In Georg Rehm and Yunyao Li, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 310–317, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-288-6. doi: 10.18653/v1/2025.acl-industry.23. URL https://aclanthology.org/2025.acl-industry.23/.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, page 274–288, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-030-99735-9. doi: 10.1007/978-3-030-99736-6_19. URL https://doi.org/10.1007/978-3-030-99736-6_19.

Kehan Long, Shasha Li, Pancheng Wang, Chenlong Bao, Jintao Tang, and Ting Wang. Recommending missed citations identified by reviewers: A new task, dataset and baselines. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13699–13711, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1196/.

João Eduardo Batista, Emil Vatai, and Mohamed Wahib. Think before you attribute: Improving the performance of llms attribution systems, 2025b. URL https://arxiv.org/abs/2505.12621.

Wenyi Huang, Zhaohui Wu, Prasenjit Mitra, and C. Lee Giles. Refseer: a citation recommendation system. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, page 371–374. IEEE Press, 2014. ISBN 9781479955695.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1149. URL https://aclanthology.org/N18-1149/.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In

Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL `https://aclanthology.org/L08-1005/`.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL `https://aclanthology.org/2020.emnlp-main.609/`.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, CSLAW '25, page 169–193, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714214. doi: 10.1145/3709025.3712219. URL `https://doi.org/10.1145/3709025.3712219`.

A Appendix

A.1 Related Work

Recent research has focused heavily on enhancing the ability of LLMs to generate correct source citations. These efforts can be broadly categorized into two groups based on the underlying paradigm they follow.

A.1.1 Generation-time Citation (`G-Cite`)

Methods: Methods in this category, which we term G-Cite, generate citations concurrently with the text. Foundational work in this area includes benchmarks like ALCE and CiteBench [Funkquist et al., 2023], which provide datasets and prompting-based baselines. Building on these, other prompting-based approaches, such as Learning to Plan [Fierro et al., 2024], further refine the attribution capabilities of LLMs. In contrast to prompting, methods like FRONT [Huang et al., 2024], LongCite [Zhang et al., 2025], and Self-Cite [Chuang et al., 2025] aim to improve source citation by fine-tuning the models themselves. Additionally, methods such as ReCLAIM [Xia et al., 2025] and Chain-of-Thought (CoT) Citation [Ji et al., 2024] build on RAG systems to enhance source citation capabilities. Other works focus on specific aspects, such as the granularity of the generated citations [Li et al., 2024b].

Datasets: In open domain short-context datasets, where the documents containing the citation to the correct answer are relatively short, include ASQA [Stelmakh et al., 2022], QAMPARI [Amouyal et al., 2023], and ALCE. For long-context datasets, where the length of the documents containing the citation is relatively large, there exist popular datasets like ELI5 [Fan et al., 2019] and LongBench-Cite.

A.1.2 Post-hoc Citation (`P-Cite`)

Methods: The second category, P-Cite, includes methods that generate citations for a pre-existing text. A prominent example is the REASONS benchmark [Saxena et al., 2025], which provides a dataset and Retrieval-Augmented Generation (RAG) baselines for sentence-level citation. Other methods also leverage RAG, using iterative approaches [Li et al., 2024a] or adding a verification step for each claim, as seen in RARR. Another recent approach that builds on top of RAG systems is CiteFix [Maheshwari et al., 2025], it uses a two step citation correction method to achieve better citation quality. Beyond standard RAG, approaches in this paradigm include fine-tuning with models like CiteBART [Çelik and Tekir, 2025] and using hierarchical attention mechanisms as in HAtten [Gu et al., 2022]. Research in P-Cite also addresses related sub-tasks, such as recommending missed citations [Long et al., 2024] or determining if a sentence requires a citation at all [Batista et al., 2025b].

Datasets: Open-domain, claim-level datasets such as FEVER, fall into this category. In addition, scientific datasets such as RefSeer [Huang et al., 2014], PeerRead (FullTextPeerRead) [Kang et al., 2018], ACL-ARC [Bird et al., 2008], SciFact (+SciFact-Open) [Wadden et al., 2020], and REASONS are `P-Cite` native. Legal datasets such as Bar Exam QA, and Housing Statutes QA [Zheng et al., 2025] also belong to this category.

A.2 Additional Implementation Details

We implement all citation variants in PyTorch with Hugging Face Transformers. ALCE inputs (both 'qa pairs' blocks and flat items) are normalized, then sentence-split with a lightweight regex and numbered to form the Context. **Answer Generation** we keep 'MAX_NEW_TOKENS' as 256, decoding uses 'temperature=0.2', 'top_p=0.95'. We parse citations with a strict regex, remap packed indices back to original sentences, and rebuild a clean **References** block (title + sentence). For RAG P/G-cite runs we add SBERT bi-encoder retrieval and a cross-encoder reranker before prompting; zero-shot variants skip retrieval.

G-Cite Zero-shot Prompt with Example

Instructions: Answer the QUESTION using ONLY the numbered Context sentences. After each answer sentence (or clause), append a citation tag exactly like <cite>[i]</cite> or <cite>[i-j]</cite> that points to the supporting Context index(es).

- Every factual claim must have at least one <cite>[i]</cite> tag.
- After the answer, output a References block that lists only the indices you cited, one per line as: [k] <Context title> — sent local_sent_index: verbatim Context sentence.
- Do NOT add explanations; return only the answer and the References block.

Example Query: *Who is the original artist of "The Sound of Silence"?*
Example Response:

```
The original artist of "The Sound of Silence" is the American duo Simon & Garfunkel <cite>[1]</cite>.
References:  [1] (The Sound of Silence) "The Sound of Silence" is a song by the American music duo
Simon & Garfunkel.
```

P-Cite Zero-shot Prompt with Example

Instructions: You will answer the QUESTION and then insert inline numeric citations [k] to that answer after the facts you assert.

- Insert [k] markers immediately after the clause(s) they support.
- The References section must contain ONLY markers that appear in the sentence.
- Do NOT include explanations; return only the answer and the References block.

Example Query: *When was the first iPhone released?*

Example Response:

```
The first iPhone was released on June 29, 2007 [1].
References:  [1] https://en.wikipedia.org/wiki/IPhone_(1st_generation)
```

A.3 Results Tables

Dataset	Method (Paradigm)	Cit. Corr. ↑	Cit. Prec. ↑	Cit. Rec. ↑	Cov. ↑	Latency ↓
ALCE	*Zero-shot (G)*	0.130	0.156	0.111	0.274	2.925
	RAG (G)	0.319	0.422	0.257	0.340	6.513
	CoT Citation (G)	0.205	0.239	0.180	0.372	17.237
	LongCite (8B) (G)	0.253	0.271	0.236	0.282	3.531
	Zero-shot (P)	0.881	1.000	0.787	0.784	4.743
	RAG (P)	0.340	0.441	0.277	0.620	9.059
	CiteBART (P)	—	—	—	—	—
	CEG (P)	0.422	0.626	0.318	0.748	6.077
LongBench-Cite	*Zero-shot (G)*	0.099	0.127	0.081	0.112	3.211
	RAG (G)	0.167	0.163	0.171	0.577	5.533
	CoT Citation (G)	0.121	0.155	0.101	0.652	17.476
	LongCite (8B) (G)	0.134	0.173	0.097	0.632	3.171
	Zero-shot (P)	0.040	0.569	0.021	0.058	3.407
	RAG (P)	0.093	0.098	0.088	0.780	5.842
	CiteBART (P)	—	—	—	—	—
	CEG (P)	0.115	0.435	0.066	0.782	9.694

Table 2: Open-domain results (ALCE and LongBench-Cite). Metrics: Citation Correctness (Corr.), Citation Precision/Recall (Prec./Rec.), Coverage (Cov.), and Latency (s). A dash (—) marks method–dataset pairs we did *not* execute due to implementation constraints (e.g., domain-locked models, unavailable code/weights).

Dataset	Method (Paradigm)	Cit. Corr. ↑	Cit. Prec. ↑	Cit. Rec. ↑	Cov. ↑	Latency ↓
REASONS	*Zero-shot (G)*	0.017	0.015	0.020	0.954	3.719
	RAG (G)	0.282	0.268	0.298	0.802	4.111
	CoT Citation (G)	0.272	0.244	0.306	0.970	9.567
	LongCite (8B) (G)	—	—	—	—	—
	Zero-shot (P)	0.029	0.027	0.032	0.946	3.535
	RAG (P)	0.272	0.269	0.276	0.814	10.083
	CiteBART (P)	0.114	0.139	0.097	0.682	—
	CEG (P)	0.259	0.241	0.280	0.989	6.628

Table 3: Scientific-domain results (REASONS). Metrics: Citation Correctness (Corr.), Citation Precision/Recall (Prec./Rec.), Coverage (Cov.), and Latency (s). G-cite runs use a constrained "rewrite with [k]" adapter over the dataset's candidate pool.

Dataset	Method (Paradigm)	Cit. Corr. ↑	Cit. Prec. ↑	Cit. Rec. ↑	Cov. ↑	Latency ↓
FEVER	*Zero-shot (G)*	0.272	0.287	0.258	0.502	4.204
	RAG (G)	0.769	0.781	0.757	0.702	3.017
	CoT Citation (G)	0.937	1.000	0.881	0.272	3.439
	LongCite (8B) (G)	—	—	—	—	—
	Zero-shot (P)	0.212	0.344	0.153	0.752	1.011
	RAG (P)	0.671	0.717	0.630	0.754	2.840
	CiteBART (P)	—	—	—	—	—
	CEG (P)	0.766	0.827	0.713	0.744	2.370

Table 4: Fact Verification-control results (FEVER). Metrics: Citation Correctness (Corr.), Citation Precision/Recall (Prec./Rec.), Coverage (Cov.), and Latency (s). FEVER is used to calibrate the evidence-agreement judge and to report supported-claim rate.