Detecting Citation Hallucinations in Large Language Model Outputs (Student Abstract)

Nipun Misra[1*†], **Vikranth Udandarao**[2†]
[1]VIT-Vellore, India
[2]IIIT-Delhi, India
nipun.misra2022@vitstudent.ac.in, vikranth22570@iiitd.ac.in

Abstract

Large Language Models (LLMs) are increasingly employed for literature reviews, academic drafting, and scholarly writing. While their fluency accelerates knowledge synthesis, they frequently produce fabricated or erroneous references, known as *citation hallucinations* (CHs). Recent studies report hallucination rates ranging from 18% in GPT-4 to over 70% in other frontier models, with domain-specific rates as high as 88% in legal contexts. Benchmarks such as CiteME further highlight the gap between LLMs (4.2–18.5% accuracy) and human annotators (69.7%), while retrieval-augmented systems like CiteAgent demonstrate partial progress. This study examines methods for automatically detecting hallucinated citations. We present a benchmark of machine-generated references labelled with three fine-grained categories (*valid*, *partially valid*, and *hallucinated*), and propose a hybrid detection pipeline combining bibliographic retrieval, fuzzy similarity, and LLM-based verification. Preliminary experiments indicate improvements over exact matching baselines. We argue that scalable, real-time citation verification is a crucial step toward developing trustworthy LLM-based scholarly assistants and generating reproducible scientific knowledge, and outline directions for multilingual and domain-specific extensions.

Introduction

Large language models (LLMs), such as GPT-4, Claude, and LLaMA, are increasingly used for academic tasks like literature surveys, drafting paper sections, and summarising prior work (Wang et al. 2024). Their fluency makes them attractive to researchers, but a persistent problem remains: *citation hallucinations*. These include fabricated papers, misattributed authors, or plausible-looking but incorrect bibliographic metadata. Such errors risk spreading misinformation, undermining trust in AI tools, and compromising the reliability of literature reviews.

Recent empirical studies show that citation hallucinations are widespread, with 55% of GPT-3.5 citations and 18% of GPT-4 citations fabricated, and even non-fabricated references often containing substantive errors (Walters and Wilder 2023). Controlled experiments reinforce this concern: ChatGPT-4o exhibits a hallucination rate of 20.0% while Gemini Advanced reaches 76.7% (Jesson et al. 2024). In high-stakes legal contexts, hallucination rates are even more alarming, ranging from 58% with GPT-4 to 88% with LLaMA-2 when asked verifiable questions about federal cases (Dahl et al. 2024). Financial and scientific domains also reveal accuracy gaps, with substantial bibliographic error rates that can distort downstream analysis (Erdem, Hassett, and Egriboyun 2025).

Benchmarking efforts, such as CiteME, quantify this performance gap: frontier LMs achieve only 4.2–18.5% accuracy, compared to 69.7% for human annotators. However, systems that augment LMs with retrieval abilities, such as CiteAgent, demonstrate promise by achieving 35.3% accuracy on CiteME, enabling autonomous search and verification (Press et al. 2024). These findings highlight that the core challenge is not only generation but also reliable integration of retrieval into LLM workflows.

Meta-analyses underscore the broader challenge: dataset documentation and benchmarking practices often lack transparency, which complicates the reproducibility and systematic evaluation of hallucinations (DeHaan et al. 2025). The persistence of citation errors across models, domains, and benchmarks motivates a rigorous investigation into scalable detection and mitigation strategies.

This motivates the following research question:

(RQ) How can hallucinated citations be detected automatically with high precision and recall?

Methodology

Our methodology combines dataset construction and a multistage detection pipeline to study and mitigate citation hallucinations in LLM outputs.

Dataset Construction

We curated a benchmark dataset to capture hallucination behaviours across domains where citation errors are prevalent (e.g., machine learning, NLP, vision, bioinformatics, and law). Literature-style prompts (e.g., "survey key works in X," "recent advances in Y") were issued to three frontier LLMs (GPT-4o, Claude 3.5 Sonnet, LLaMA-4 Maverick). From their responses, we extracted citations (au-

*Primary Author
†These authors contributed equally.

thor, title, year, venue) and verified them against bibliographic databases including Crossref, OpenAlex, and Semantic Scholar. Each entry was labelled as:

- *Valid* — all major fields correct,
- *Partially Valid* — some overlap but one or more errors,
- *Hallucinated* — no credible match found.

Two annotators labelled all items, resolving disagreements through adjudication (Cohen 's κ reported). The pilot dataset comprises $\sim$200 citations, with expansion planned beyond 1,000.

Detection Pipeline

To automatically detect hallucinations, we designed a three-stage pipeline (Figure 1), where each stage progressively handles more ambiguous cases:

1. **Exact Lookup**: Direct matching of title + first author against bibliographic databases. If matched confidently, label as *Valid*.
2. **Fuzzy Search**: For unmatched cases, apply string similarity (Levenshtein, Jaro–Winkler), BM25 retrieval, and embedding-based similarity to retrieve candidate references.
3. **LLM-Assisted Verification:** A lightweight LLM compares each citation with top-k candidates, checking overlap in title, authors, and year to classify as *Valid*, *Partially Valid*, or *Hallucinated*, producing confidence scores. We use the *Llama-3-8B-Instruct* model for this stage, chosen for its strong instruction-following ability, factual reasoning performance, and moderate computational footprint, which is suitable for scalable citation verification.

This hybrid design enhances both precision and recall by uniting deterministic lookup with flexible retrieval and semantic verification. To foster reproducibility, the implementation and annotated dataset are publicly available on GitHub.[1]

Preliminary Experiments

We conducted a pilot evaluation on 200 citations (approximately 50 per model). Manual verification against bibliographic databases provided the following observations:

- **Hallucination Rates.** GPT-4o ($\sim$12%), Claude 3.5 Sonnet ($\sim$16%), and LLaMA-4 Maverick ($\sim$21%).
- **Exact Lookup.** Detected $\sim$65% of hallucinations but missed several partially correct cases.
- **Fuzzy Retrieval.** Increased recall to $\sim$75%, though some false positives arose from similar titles and author combinations.
- **LLM-Assisted Verification.** Combined with retrieval, achieved $\sim$80% precision in detecting hallucinated citations, representing a 15–20% relative improvement over database-only baselines.

These findings suggest that the proposed hybrid pipeline substantially enhances both precision and recall, validating the feasibility of automated citation hallucination detection.

[1]https://github.com/Vikranth3140/Citation-Hallucination-Detection

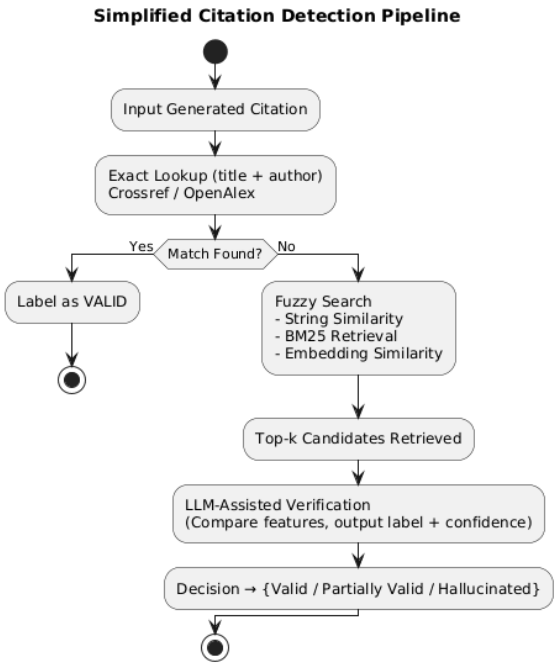

Figure 1: Multistage detection pipeline for citation verification.

Conclusion and Future Work

We presented a benchmark dataset and a hybrid detection pipeline for identifying citation hallucinations in LLM-generated academic text. By combining database lookups, fuzzy retrieval, and LLM-assisted verification, our approach achieved over 80% precision in preliminary experiments and improved recall relative to database-only baselines. These results demonstrate the feasibility of scalable, automated citation verification and its potential to enhance the reliability of LLM-based scholarly writing. In future work, we aim to scale the dataset to thousands of citations across additional domains, extend verification to multilingual settings, and publicly release both the dataset and software to promote transparency, reproducibility, and responsible use of AI in research.

Acknowledgments

We thank the reviewers for their valuable feedback, which significantly contributed to the improvement of the content and clarity of this work. We also acknowledge the use of the Crossref, OpenAlex, and Semantic Scholar APIs for bibliographic data retrieval and verification, as well as the Ollama platform for providing API access to the Llama-3-8B-Instruct model used in this study.

References

Dahl, M.; Magesh, V.; Suzgun, M.; and Ho, D. E. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1): 64–93.

DeHaan, S.; Liu, Y.; Bollen, J.; and Blanco, S. A. 2025. GPT Editors, Not Authors: The Stylistic Footprint of LLMs in Academic Preprints. arXiv:2505.17327.

Erdem, O.; Hassett, K.; and Egriboyun, F. 2025. Hallucination in AI-generated financial literature reviews: evaluating bibliographic accuracy. *International Journal of Data Science and Analytics*.

Jesson, A.; Beltran-Velez, N.; Chu, Q.; Karlekar, S.; Kossen, J.; Gal, Y.; Cunningham, J. P.; and Blei, D. 2024. Estimating the Hallucination Rate of Generative AI. arXiv:2406.07457.

Press, O.; Hochlehnert, A.; Prabhu, A.; Udandarao, V.; and Bethge, M. 2024. CiteME: Can Language Models Accurately Cite Scientific Claims? In *Proceedings of the NeurIPS Datasets and Benchmarks Track*.

Walters, W. H.; and Wilder, E. I. 2023. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1): 14045.

Wang, J.; Hu, H.; Wang, Z.; Yan, S.; Sheng, Y.; and He, D. 2024. Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-stage Scholars. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.