RefLens: End-to-End Evidence-Grounded Citation Verification with LLM Agents

SeungHoo Lee[1], **JuneHyoung Kwon**[2], **Jooweon Choi**[2], **JungMin Yun**[2], **Seunguk Yu**[2], **Yoonji Lee**[2], **Jinhee Jang**[2], **YoungBin Kim**[1,2]

[1]Graduate School of Advanced Imaging Sciences, Multimedia and Film, Chung-Ang University
[2]Department of Artificial Intelligence, Chung-Ang University
{hahaha1321, dirchdmltnv, jwjwchoi0910, cocoro357, bokju128, pioneer0305, jinheejang, ybkim85}@cau.ac.kr

Abstract

Accurate citation is critical, yet error rates remain high across scientific literature. We present RefLens, an end-to-end system that automates citation verification from PDF parsing to interactive report generation. Unlike summary- or embedding-based approaches, RefLens performs evidence-grounded verification by extracting verbatim spans from original sources and displaying citation-level cards and a paper-level dashboard. In a 35-participant study, users rated value (M=4.34), trust (M=4.15), and usability (M=4.19) highly, with strong adoption intention (M=4.28).

Project Page — https://hoohahabighead.github.io/RefLens
Code —
https://github.com/hoohahaBIGHEAD/RefLens-app

Introduction & Related Work

Accurate citation is a critical component of the reliability of academic research, yet 11-41% of published papers contain reference errors, severely undermining research reproducibility (Zhang and Abernethy 2024). The recent advances of large language models (LLMs) has exacerbated this issue by introducing *citation hallucination*—the generation of plausible but non-existent sources—which further burdens already overloaded reviewers (Drozdz and Ladomery 2024).

While the field has evolved toward semantic verification of scientific claims (Wadden et al. 2020), existing interactive systems still primarily focus on bibliographic management or summary-based analysis, which can leave readers and reviewers exposed to hallucination risks (Chelli et al. 2024). To date, a true end-to-end automated system that performs evidence-based semantic verification by directly acquiring original source documents has been absent. To overcome these limitations, we propose RefLens, a multi-agent LLM framework that performs verification based on direct quotes from source documents. RefLens aims to enhance the transparency and reliability of scholarly arguments by automating the entire verification process and fundamentally reducing the hallucination risks inherent in summarized information.

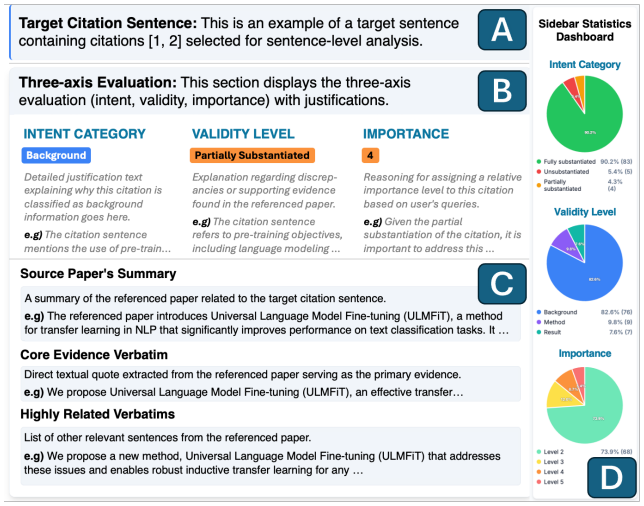

Figure 1: Example of a citation verification card in RefLens. [A] shows the target citation sentence. [B] displays the three-axis evaluation (intent, validity, importance) with justifications. [C] presents the source paper's summary and the core evidence. The sidebar [D] shows the overall citation statistics dashboard for the entire paper.

Method: RefLens

RefLens automates the entire citation verification process through a modular pipeline architecture, adopting principles from recent surveys on LLM-based autonomous agents (Wang et al. 2023), as depicted in Figure 2, where multiple autonomous agents collaborate. The workflow comprises a preprocessing phase to acquire and structure documents, followed by a hierarchical verification phase for analysis, culminating in an interactive report.

Preprocessing: Document Acquisition. This initial phase builds the foundational data for verification. A **PDF Processor** agent analyzes the target paper's layout, converting the PDF into structured text while precisely extracting in-text citations and the bibliography. Leveraging this bibliographic information (e.g., DOI, arXiv ID), a **Crawler** agent then automatically downloads the original source PDFs from online academic databases, creating a comprehensive evidence database.

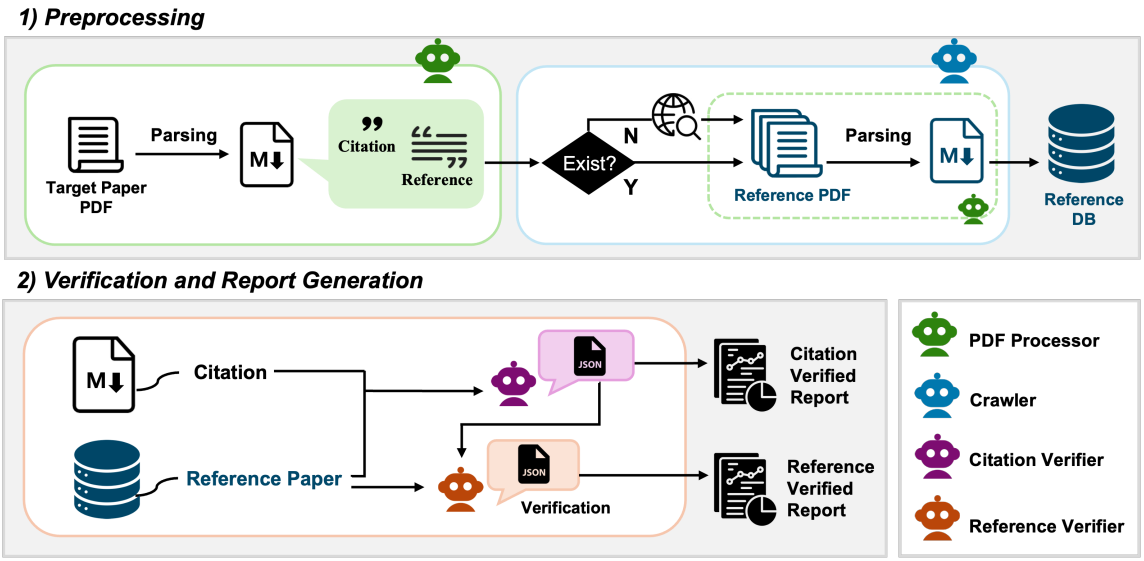

Figure 2: RefLens system pipeline overview. The framework operates through two main phases: (1) **Preprocessing** where the PDF Processor extracts citations and references from the target paper, and the Crawler automatically acquires missing reference documents to build a comprehensive database, and (2) **Verification and Report Generation** where the Citation Verifier and Reference Verifier analyze citation-source pairs to produce dual complementary reports for comprehensive citation validation.

Hierarchical Verification: Evidence-Grounded Analysis. This is the core phase where verification is performed using the constructed evidence database. The **Citation Verifier** agent conducts a microscopic analysis—defined as a fine-grained evaluation of a single citation-source pair—based on a three-axis evaluation model: intent (Cohan et al. 2019), validity (Zhang and Abernethy 2024), and user-defined importance. The key technology is a **Prompt-based Evidence Extraction** methodology. RefLens does not use common retrieval-augmented generation techniques or vector embeddings. Instead, it directly leverages the context window length of state-of-the-art LLMs. It provides the model with the entire structured text of the original reference and the target citation in a single prompt, instructing it to identify and extract verbatim the most relevant text spans that support or contradict the claim. This approach ensures that all verification decisions are anchored in the actual source text, fundamentally reducing hallucination risks. If extracted evidence is invalid or non-existent, it is added as a negative case example, incorporating a feedback loop that enhances accuracy. The results of this analysis are presented to the user in an interactive verification card, as shown in Figure 1. Finally, a **Reference Verifier** agent aggregates these micro-level results to perform a macroscopic, reference-level analysis and generate a strategic reading guide for the final report.

User Studies & Evaluations

To evaluate the practical utility of RefLens, we conducted a user study with 35 participants from diverse academic backgrounds (students and researchers). After experiencing a system demo, participants rated the system on a 5-point Likert scale across four dimensions. As shown in Table 1, the results indicate the system's high practical utility. Users rated its overall value and impact (4 questions, M=4.34), trustworthiness (4 questions, M=4.15), and usability (4 questions, M=4.19) highly. Strong adoption intention (3 questions, M=4.28) further validates that our evidence-grounded approach effectively addresses real-world research challenges. Details and limitations of the study are provided on our project page.

Dimension	Mean (SD)	Cronbach's α
Overall Value and Impact	4.34 (0.84)	0.89
Information Quality & Trust	4.15 (0.80)	0.82
System Usability	4.19 (0.77)	0.85
Adoption Intention	4.28 (0.89)	0.87

Table 1. User evaluation results for key dimensions (N=35).

Conclusions & Future Work

This paper introduced RefLens, an end-to-end, evidence-grounded framework for automated citation verification. We presented a novel methodology that utilizes prompt engineering instead of vector embeddings to extract evidence directly from source documents, and a user study with 35 participants demonstrated its strong utility and trustworthiness. This work not only provides a practical tool but also suggests a new paradigm for human-AI collaboration in ensuring scholarly integrity. By grounding LLM-generated analysis in verifiable evidence, RefLens demonstrates a path toward building more trustworthy and accountable AI systems for critical domains. Future work will focus on broader accuracy benchmarking, pipeline robustness, expanded user studies, and multilingual support. The system demo is available on our project page.

Ethical Statement

Our work focuses on assisting researchers and reviewers in checking the consistency of citations. The user study involved 35 adult volunteers. All participants provided informed consent, and all responses were fully anonymized, following our institution's ethical guidelines for low-risk human-subject research. Participants received a small token of appreciation for their time.

RefLens is designed to operate exclusively on documents that users are already permitted to access. The system's Crawler module attempts to retrieve documents from publicly accessible sources (e.g., open-access repositories) or materials available through the user's institutional subscriptions; it does not bypass paywalls. In cases where the automated Crawler cannot retrieve a source, users may manually add the document to the system's local database. This process relies on the user's own resources on the premise that the user has obtained these documents through legitimate means. The system does not redistribute or republish the full text of any source document, respecting copyright and publisher licenses.

We acknowledge the potential for misuse (e.g., superficially "correcting" low-quality, AI-generated text without critical understanding). RefLens is intended as a decision-support tool to complement, not replace, human judgment. Its reports are designed to assist good-faith scholarly review, and we caution against relying on it as a fully automated verification system, as it remains a supplementary tool.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00556246).

References

Chelli, M.; Descamps, J.; Lavoué, V.; Wintenberger, N.; and Morel, G. 2024. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26: e53164.

Cohan, A.; Ammar, W.; Van Zuylen, M.; and Cady, F. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3586–3596.

Drozdz, J. A.; and Ladomery, M. R. 2024. The peer review process: Past, present, and future. *British Journal of Biomedical Science*, 81.

Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; Van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7534–7550.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.-Y.; Tang, J.; Chen, X.; Lin, Y.; et al. 2023. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Zhang, T. M.; and Abernethy, N. F. 2024. Detecting reference errors in scientific literature with large language models. *arXiv preprint arXiv:2411.06101*.