Detecting AI-Generated Fabricated Reference with an Automated Reference Verification System

1. Introduction

Recently, generative artificial intelligence (Generative AI) empowered by large language models (LLMs) has been widely adopted in academic writing and text refinement. However, these AI tools frequently produce fabricated references, including pseudo-references, mismatched pairings of real authors with fabricated articles, and erroneous or unresolvable DOI numbers that do not correspond to any actual publication. Such fabricated references pose substantial risks to the quality of scholarly communication and research reproducibility(van Rensburg, 2025). As students and researchers increasingly rely on generative AI for literature gathering and drafting, AI-generated fabricated references not only raise the verification burden for editors and reviewers but also present significant challenges to academic integrity and assessment systems (Haan, 2025). Empirical evidence further shows that, within AI-assisted systematic reviews, the rates of citation errors and fabricated references are notably high, requiring authors to conduct rigorous manual validation to ensure the reliability of academic data (Chelli et al., 2024).

Under this background, this study investigates a core question: whether an automated verification mechanism can effectively identify fabricated references in academic papers without increasing manual workload. Furthermore, this study evaluates the system's performance in terms of detection accuracy and practical usability to assess its feasibility as a supporting tool for journals, conferences, and higher education contexts.

The primary objectives of this study are as follows: First, to design and implement an automated system for verifying references in English academic papers and flagging potentially fabricated references. Second, to evaluate the system's effectiveness in detecting AI-generated fabricated references through empirical testing. Third, to examine the system's potential applications in academic review processes and academic integrity from both scholarly and administrative perspectives.

From an academic standpoint, this study addresses growing concerns regarding AI-generated fabricated references, which are outcomes of AI hallucinations (Adel & Alani, 2025). These AI-generated fabricated citations are not only mistakes that need to be corrected, but they are also signals that parts of the academic articles are generated by AI and not carefully confirmed by human authors (Bender et al., 2021).

From a practical perspective, the proposed automated reference verification system provides editors, reviewers, authors, advisors, and graduate students with an operational automated support mechanism to mitigate the risk of fictitious references entering the academic ecosystem.

From an academic perspective, the detection of fabricated references is a screening indicator for AI hallucination issues(Ji et al., 2023). Fabricated references reflect risks to academic integrity and align with recent empirical observations of AI-generated hallucinated content(Maynez et al., 2020). Although the proposed automated reference verification system cannot detect all types of AI-generated hallucinations in academic papers, the reference check results can substantially narrow the scope of manual checking, enabling editors and reviewers to concentrate their limited time on high-risk references. Additionally, the automated reference verification system provides researchers with a self-checking tool, reducing the likelihood of inadvertently citing fabricated references(van Dis et al., 2023).

The study aims to develop a system that checks references using multiple scholarly databases, incorporating reference parsing and similarity comparison techniques, to effectively enhance the detection of fabricated references. The system offers benefits in addressing the academic integrity challenges raised by generative AI. By using the developed systems, the study checks for the existence of fabricated references included in students' theses. The results of the reference check can report the existence of fabricated literature in student theses, which may serve as a cue for the presence of AI-generated hallucinated content.

2. System Development

2.1 Objectives of System Development

The automated reference verification system developed in this study is primarily intended to address academic integrity challenges arising from the widespread adoption of generative AI, particularly the precise detection of fabricated citations produced through AI hallucinations. A core operational requirement of the system is the capability to directly process raw reference lists input by the user. By utilizing a robust text parsing module, the system transforms unstructured citation strings into structured bibliographic fields, such as author names, titles, publication years, and DOIs. To ensure the authority and reliability of verification results, the system integrates APIs from multiple major scholarly databases, including Crossref, Scopus, OpenAlex, and Google Scholar, in order to query and confirm the authenticity of cited

references. Furthermore, to enhance practical applicability, the system is designed to support automated cross-checking processes that reduce the time cost and omission risks associated with manual verification by editors and researchers. In addition to basic existence checks, the system performs real-time connectivity tests on DOIs and URLs to ensure that references not only exist within bibliographic databases but also provide valid and accessible online pathways. Finally, the system generates structured diagnostic reports and visualized statistical summaries that explicitly indicate specific types of citation errors, thereby serving as an effective tool for academic review and quality control.

2.2 System Environment

To balance cross-platform deployability, development efficiency, and long-term maintainability, the system architecture in this study adopts a multi-language and modular integration design. The core backend logic is primarily implemented in Python, leveraging its mature and extensive ecosystem to support key functionalities such as citation parsing, data cleaning, fuzzy matching, and result filtering. Python's comprehensive text-processing libraries and regular expression mechanisms enable effective pattern recognition and field extraction from unstructured academic texts, thereby enhancing the stability and accuracy of citation structuring.

The system's front-end interactive interface is developed using the Streamlit framework, which allows backend Python computational workflows to be rapidly transformed into interactive web-based applications. This design lowers the barrier to use and improves overall system usability. Through this interface, users can input reference lists for verification and immediately review parsing and validation results, thereby supporting real-time decision-making and iterative revision during the research process.

At the citation parsing layer, this study integrates the AnyStyle engine, which operates within a Ruby runtime environment, as the primary tool for transforming unstructured references into structured data. AnyStyle demonstrates strong adaptability to diverse citation formats and supports the parsing of bilingual Chinese–English references, effectively reducing parsing failures caused by format inconsistencies. The parsed outputs are subsequently passed to downstream modules in standardized field formats, ensuring consistency and reliability throughout the system's data flow.

In terms of data access and performance optimization, the system employs the Pandas library to manage local CSV-based literature databases, enabling efficient searching and similarity matching operations. Additionally, to support verification workflows that require querying multiple external academic APIs, the system adopts a multithreading mechanism to parallelize request processing. This approach effectively reduces overall response latency and significantly improves processing throughput in large-scale reference verification scenarios.

The system environment of this study is illustrated in Figure 1.

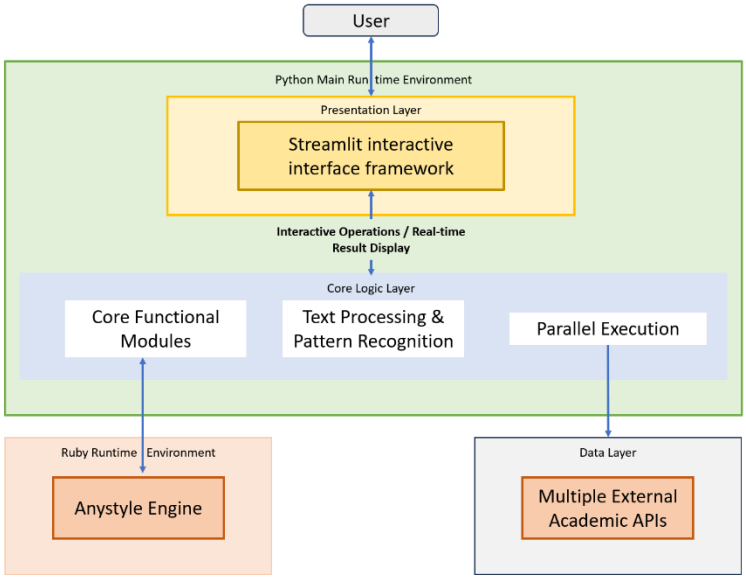

Figure 1:System Environment

2.3 System Design and Architecture

The system proposed in this study adopts a modular architectural design. The overall workflow can be mainly divided into three core modules: Text Parsing, Multi-source Verification, and Report Generation. These modules are loosely coupled and interconnected through structured data formats, thereby enhancing system extensibility, maintainability, and future scalability.

During the text parsing stage, the system first receives unstructured citation text provided by the user. To mitigate parsing difficulties caused by variations in citation

styles and language differences, the system employs the AnyStyle engine as the primary parsing tool. AnyStyle performs citation line detection and language identification, converting raw textual references into structured JSON objects. Each citation is standardized into key bibliographic fields, including authors, title, container title (e.g., journal or conference name), and publication year. Given that AI-generated content often exhibits irregular punctuation, duplicated parentheses, or non-compliant citation formats, additional field-cleaning and normalization procedures are implemented at this stage. These procedures automatically correct abnormal symbols and formatting inconsistencies, ensuring data consistency and reliability for subsequent verification processes.

Following text parsing, the system proceeds to the multi-source verification stage. For each structured citation, the system sequentially queries major scholarly databases—such as Crossref, Scopus, and Google Scholar—to determine the actual existence of the cited work and to retrieve authoritative identifiers, including DOIs or official landing page URLs. Furthermore, when explicit web links are present within a citation, the system performs URL accessibility checks by evaluating HTTP response statuses, thereby filtering out invalid links or fabricated paths that point to non-existent resources.

In the report generation stage, the system consolidates parsing and verification results and categorizes each reference according to its verification status. The outcomes are transformed into structured outputs suitable for further analysis. A visualized statistical dashboard presents the overall distribution of references—for example, verified references, partially verifiable references, and high-risk references—enabling users to quickly assess citation quality at a glance. In addition, the system provides downloadable CSV files containing complete verification results, facilitating subsequent manual review or integration into other research workflows. Through this design, the system supports early-stage identification of potential AI-generated hallucinated citations during the manuscript preparation process, thereby contributing to automated academic integrity assurance.

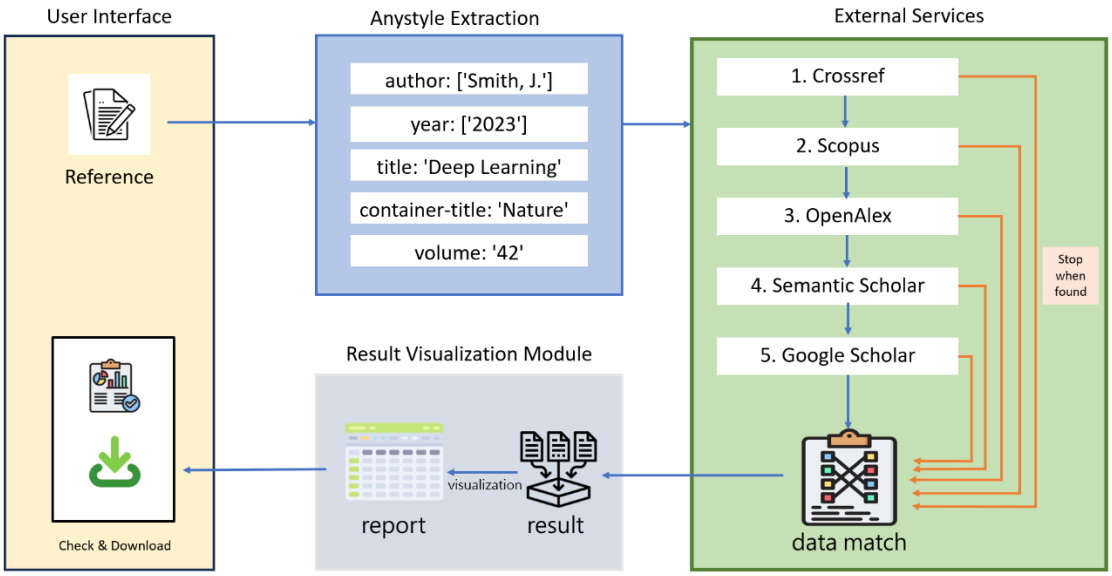

Figure 2:System Architecture

2.4 Module Design

To achieve the research objective of cross-database automated citation verification, the proposed system is structured into five interconnected core modules, each responsible for a distinct functional stage: document parsing, reference extraction, citation verification, citation style detection, and result visualization. This modular design reduces overall system complexity while enabling independent testing, refinement, and extension of individual components, thereby enhancing system maintainability and scalability.

(1) Reference Parsing Module

The Reference Parsing Module converts raw citation text into structured data using a high-performance AnyStyle engine. To accommodate the complexity of multilingual manuscripts, this module features a Language-Adaptive Parsing mechanism: The system automatically detects Chinese characters via Unicode range scanning for each citation line. If CJK characters are identified, the module invokes a customized parsing model (custom.mod) to handle non-Western bibliographic structures; otherwise, it employs the standard AnyStyle default model. Following the parsing process, a Normalization Layer applies Unicode NFKC normalization and heuristic cleaning to eliminate formatting noise (e.g., special symbols and layout

inconsistencies). This ensures that critical metadata—such as authors, titles, and DOIs—is accurately extracted, providing a standardized data foundation for subsequent database verification..

(2) Citation Verification Module

The Citation Verification Module constitutes the core functionality of the proposed system and is responsible for cross-database validation of citation authenticity. By integrating Crossef API, OpenAlex API, Semantic Scholar API, Scopus API and Google Scholar API,the system queries bibliographic records using citation titles and author information, and computes similarity scores between the input citations and retrieved database entries. Through multi-source cross-validation, the system effectively identifies non-existent publications, mismatches between authors and titles, and fabricated references generated through AI hallucinations.

(3) Result Visualization Module

The Result Visualization Module utilizes Streamlit as the front-end environment to present verification outcomes in an intuitive and user-friendly manner. The interface displays citation verification status categories, summaries of detected error types, and detailed information for references flagged as high-risk. Through visualized statistics and grouped presentation, users can rapidly assess the overall quality of citations and focus their attention on entries requiring further manual inspection, thereby improving the efficiency of academic review and self-checking processes.

3. Evaluation

3.1 Performance Evaluation Metrics

This study aims to quantitatively evaluate the performance of the reference checking system in distinguishing between real references and fake references. The core objective of the evaluation is to assess whether the system can correctly verify existing citations and accurately identify non-existent or non-retrievable references. To ensure an objective analysis of system performance, this section defines the four quadrants of the confusion matrix and employs Precision, Recall, F1-score and

Accuracy as quantitative evaluation metrics.

3.1.1 Definition of the Confusion Matrix

In this experiment, fake or hallucinated references are defined as positive class, representing the primary detection target of the system, while real references are defined as negative class.

The definitions of each quadrant of the confusion matrix are as follows:

- True Positive (TP):
  The input is a fake reference, and the system correctly reports it as “not found” (successful detection).
- False Positive (FP):
  The input is a real reference, but the system incorrectly reports it as “not found” due to database coverage limitations or parsing errors (misclassifying a real reference as fake).
- True Negative (TN):
  The input is a real reference, and the system successfully retrieves a matching record from the database and reports it as “verified.”
- False Negative (FN):
  The input is a fake reference, but the system incorrectly matches it to an unrelated record and reports it as “verified” (failure to detect a fake reference, resulting in a missed detection).

3.1.2 Evaluation Metrics

Given the task-specific characteristics of fake reference detection, this study adopts the following statistical metrics, defined as follows:

- **Precision:**
  Precision reflects the proportion of references identified by the system as “fake” that are indeed fake. A higher Precision indicates a lower false positive rate, meaning fewer real references are incorrectly flagged as fake.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:**

Recall reflects the proportion of actual fake references that are successfully detected by the system. Recall serves as a key indicator of system robustness, with higher values indicating fewer fake references escaping detection.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score:**

  The F1-score is the harmonic mean of Precision and Recall. Since both excessive false positives (low Precision) and excessive false negatives (low Recall) can undermine user trust in a reference checking system, the F1-score is adopted as a comprehensive metric to evaluate the system's balanced performance.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Accuracy**

  Accuracy is included as a supplementary evaluation measure to reflect the overall classification performance of the system. Accuracy represents the proportion of correctly classified instances among all evaluated references, including both fake and real references.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

3.1.3 Experimental Results and Confusion Matrix

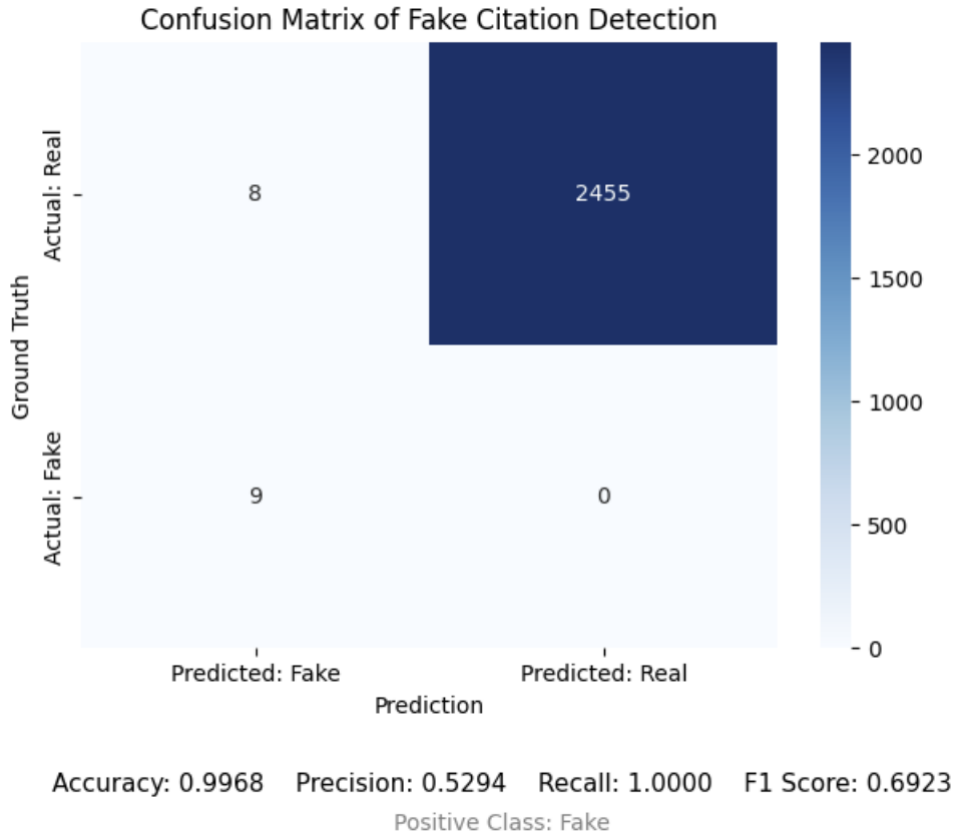

Figure 3:Confusion Matrix

To evaluate the practical performance of the proposed system in real academic scenarios, this study selected 32 doctoral dissertations in Information Management from academic years 113 to 114 in the Taiwanese thesis database as the test sample. After excluding 5 dissertations without reference lists, a total of 2,472 citations were included for testing. The test data encompassed three categories: "real and correctly formatted," "partially erroneous," and "completely non-existent (hallucinated)." Prior to system evaluation, the authenticity and category of all samples were pre-verified through the thesis database to establish the experimental ground truth. Subsequently, the data was imported into the system for automated detection, with a focus on assessing the system's capability to identify anomalous citations. Items flagged by the system as potentially erroneous underwent manual re-verification (e.g., individual DOI or Google Scholar searches) to validate judgment accuracy.

Figure 3 presents the confusion matrix results for this test set, which contains 9 manually verified fake citations alongside the remaining real citations. The matrix shows the system successfully marked all 9 fake citations as "not found" (TP = 9, FN = 0) and classified 2,455 citations as matching database records (TN = 2,455).

However, the system misclassified 8 real citations as fake (FP = 8), attributable to two primary causes: (1) inability to retrieve authentic document URLs via title matching alone, such as "J. Redmon and A. Farhadi, 'YOLO v.3,' Tech Rep., pp. 1–6, 2018," where the research team obtained the original via cross-referencing other papers; and (2) citations hosted in sources outside the system's database coverage (e.g., thesis repositories). These results yield an overall accuracy of 99.68%, fake citation precision of 52.94%, recall of 100.00%, and F1-score of 69.23%, demonstrating complete detection of all fake citations while indicating room for improvement in true citation identification, addressable through refined matching logic to reduce false positives.

In summary, despite the system's currently focused scope, these results from 2,472 citations confirm its ability to systematically address hallucinated citation challenges without substantially altering existing workflows. For journal editors and educational settings, the system offers a more efficient citation review mechanism that enhances problem detection probability. For future research, these preliminary findings validate citation-level verification as a practically feasible foundation, paving the way for content-level validation upon accumulating more data and optimizing false positive issues.

4. Application

4.1 The Urgency and Challenges of Practical Application

With the rapid proliferation of Generative AI technology, the academic community faces unprecedented challenges. While Large Language Models (LLMs) have significantly improved the efficiency of research and writing, the accompanying phenomenon of hallucination has led to the generation of a substantial volume of fabricated literature. These AI-generated citations often appear rigorous in format and reasonable in content, yet they are entirely non-existent. This phenomenon has severely contaminated the authenticity of academic literature and has become one of the most destructive threats to academic integrity today.

Faced with this rapidly worsening problem, current review mechanisms are increasingly inadequate. Traditional reference verification primarily relies on manual review by reviewers or editors. However, this task is extremely time-consuming and labor-intensive. Given the surge in submission volumes and AI-generated content,

relying solely on human effort to verify databases individually has become impractical. Furthermore, manual verification is susceptible to errors caused by fatigue and cognitive bias, allowing false information to infiltrate formal academic records.

If this trend is not effectively curbed, the proliferation of fabricated citations will severely erode the foundation of trust in academic research. This could lead to subsequent studies being built upon erroneous evidence, which not only wastes scientific resources but also causes irreversible damage to academic ethics. Therefore, developing a detection system capable of automated, high-efficiency, and precise identification of fake citations is no longer merely a matter of technical optimization; it is an urgent necessity for maintaining a healthy academic ecosystem.

4.2 Workflow and Integration

The system proposed in this study aims to elevate citation checking from a labor-intensive manual activity to a standardized step embedded within the overall workflow. Specifically, users can input a list of references, which the system structures into standardized representations containing fields such as author, title, journal name, year, volume, issue, and DOI. Subsequently, the system connects simultaneously to multiple data sources, including Crossref, Scopus, OpenAlex, and Google Scholar, to attempt to find a corresponding physical entity for each citation. Based on the matching results, citations are marked as "Verified" or "Requires Manual Review." Finally, a review report specifically targeting citations is generated to assist editors and reviewers in their decision-making process.

Regarding workflow integration, the system adopts a user-interface-centric design intended to allow editors and reviewers to incorporate it into their daily operations with a minimal learning curve. Editors simply need to import the manuscript file or reference list into the system while reviewing the paper. On a single interface, they can view a summary of citation risks generated by the system, allowing them to quickly judge whether the manuscript contains a high proportion of suspicious literature. They can then decide whether to request clarification from the author or conduct further manual checks. At the operational level, the system maintains a streamlined interface and process, enabling users to complete citation checks in a very short time without altering their original submission or review habits. By simply

uploading the manuscript or its reference list, the system automatically organizes clear inspection results and warning markers. This allows users to identify citation items worthy of further attention at a glance, thereby reducing the cost of adoption while effectively enhancing efficiency and confidence when dealing with hallucinated citations.

4.3 Targeted Application Scenarios

For editors who need to handle a certain volume of manuscripts, verifying citations individually within a limited time is often impractical. Consequently, they mostly rely on experience and random sampling, which inevitably leads to oversights. Through this system, users can quickly view a structured citation list and basic inspection results after importing the references. This makes it easier to spot obviously suspicious items or inconsistencies in format and content, thereby determining whether to ask the author for an explanation or to proceed with a more detailed manual verification.

In the fields of higher education and research training, this system also responds to new challenges faced by instructors in recent years. As students increasingly use Generative AI to write class reports and theses, the concerns of advisors and committee members regarding text credibility are no longer limited to plagiarism and similarity rates. Instead, they now extend to whether the bibliographical basis itself truly exists and whether it has been correctly interpreted and cited. Faced with dozens of reports in a class or theses in a department, checking every reference individually is nearly impossible. In this context, this system serves as a screening tool for instructors and administrators. It allows for the batch scanning of assignments and theses, marking works with abnormally high proportions of hallucinated citations or concentrations of highly suspicious entries. This enables instructors to focus their limited time and energy on a few high-risk cases and to engage in deeper dialogue regarding academic ethics and literature usage during interviews or oral defenses.

Beyond academic publishing, this system is applicable to other text writing and review contexts that rely on literature as references, such as medical and health-related reports, policy drafts, or industrial research documents. Although these texts differ in nature from academic papers, they similarly involve the citation and interpretation of research results. Once cited content contains obvious errors or

questionable sources, it may affect subsequent discussions and judgments. By processing citations in drafts through this system for organization and preliminary checking, reviewers can browse citation lists more systematically. They can confirm parts that appear unusual or inconsistent with the document's theme, thereby slightly reducing the risk of misunderstanding caused by citation distortion without adding excessive burden.

5. Discussion & Conclusion

The automated academic reference verification system developed in this study demonstrated high classification accuracy (Accuracy: 99.68%) and strong risk interception capability in empirical evaluations. Given the heterogeneity of references in academic manuscripts, including journal articles, conference papers, technical reports, and web-based resources. The core contribution of this system lies in its dual-track verification architecture. Standard references with DOIs are verified through precise identifier matching, while non-standard sources lacking formal indexing are validated via cross-database retrieval and real-time availability checks.

A key experimental outcome is the achievement of a perfect recall rate (Recall: 1.0000), indicating that all hallucinated references in the evaluated dataset were successfully detected. Although a limited number of real references were incorrectly flagged as fake, resulting in a moderate precision value (Precision: 0.5294), this reflects a design choice that prioritizes conservative detection under an academic integrity–oriented framework. By flagging ambiguous cases for further manual review, the system effectively supports a “system-assisted screening with human validation” workflow, enabling reviewers to focus on high-risk references.

With respect to future research directions, this study suggests extending the proposed framework toward cross-lingual knowledge graph integration and exploring the feasibility of embedding the system into journal submission platforms or institutional review systems. By incorporating automated pre-screening at early stages of the academic lifecycle, defensive mechanisms can be established at the source. The dual-track verification pathway and hierarchical search strategy proposed in this study not only provide an immediate solution for current citation verification challenges but also lay a solid foundation for the development of a more intelligent scholarly ecosystem. Institutionalizing such technical tools within academic workflows

represents an effective approach to maintaining academic integrity in the digital era.

References

Adel, A., & Alani, N. (2025). Can generative AI reliably synthesise literature? exploring hallucination issues in ChatGPT. *AI & SOCIETY*, *40*(8), 6799–6812. https://doi.org/10.1007/s00146-025-02406-7

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ?* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada. https://doi.org/10.1145/3442188.3445922

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis [Original Paper]. *J Med Internet Res*, *26*, e53164. https://doi.org/10.2196/53164

Haan, S. (2025). SemanticCite: Citation Verification with AI-Powered Full-Text Analysis and Evidence-Based Reasoning. *arXiv preprint arXiv:2511.16198*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, *55*(12), Article 248. https://doi.org/10.1145/3571730

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020, July). On Faithfulness and Factuality in Abstractive Summarization. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* Online.

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*, *614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

van Rensburg, L. (2025). AI-Powered Citation Auditing: A Zero-Assumption Protocol for Systematic Reference Verification in Academic Research. *arXiv preprint arXiv:2511.04683*.