

Automating Reference Integrity: An Empirical Evaluation of Detecting Hallucinated Citations in Scholarly Manuscripts with Citcheck

Type of article: Original Research

Abdeldjalil Khelassi

Associate Professor, Abou Bekr Belkaid University of Tlemcen, Algeria

General Manager, Knowledge Kingdom Publishing

Abstract

Introduction: The widespread adoption of Large Language Models (LLMs) in academic writing has introduced significant risks to scholarly publishing through the generation of hallucinated, non-existent references ("ghost citations"). Manual verification of citation integrity is labor-intensive and error-prone for editorial teams. This study evaluates an automated computer-implemented system, called: **Citcheck**, designed to analyze manuscript bibliographies and verify citation network integrity.

Methods: An empirical benchmark dataset ($N = 147$) was evaluated across three distinct groups: Real Indexed Literature ($n = 47$), Real Edge-Case Literature ($n = 50$), and Synthetic Ghost Citations ($n = 50$). Citcheck parsed unstructured bibliographic strings, queried the Crossref database, and calculated fuzzy similarity metrics for titles, authors, and publication years. Works falling below defined threshold criteria were flagged as "Conflict / Low Match." Performance was assessed using sensitivity, specificity, precision, accuracy, and Matthews Correlation Coefficient (MCC).

Results: The system correctly identified and blocked 47 out of 50 hallucinated citations, achieving a **Specificity of 94.00%** and an overall **Precision of 95.59%** (False Positive Rate = 6.00%, $n = 3$). For real literature, the platform successfully verified 65 out of 97 entries, yielding a **Sensitivity of 67.01%** (False Negative Rate = 32.99%, $n = 32$). Across the entire sample ($N = 147$), the automated system demonstrated an overall **Accuracy of 76.19%** and an **MCC of 0.58**.

Discussion: The high specificity and precision confirm that automated API-driven fuzzy matching serves as a highly effective first line of defense against synthetic AI-generated citations in editorial workflows. False negatives primarily occurred among non-standard edge-case literature (e.g., niche monographs, institutional reports, and preprints lacking uniform metadata formatting). Because deterministic algorithms cannot reliably distinguish non-indexed genuine publications from well-structured hallucinations, the system is designed under a Human-in-the-Loop (HITL) governance model. Machine automation accelerates initial screening and flags anomalies, while human managing editors retain final adjudicative oversight for edge-case citations.

Keywords: Citation Integrity, Ghost Citations, Large Language Models, Bibliometrics, Automated Verification.

Corresponding author: Abdeldjalil KHELASSI, Abou Bekr Belkaid University of Tlemcen, Algeria.

Email: khelassi.a@gmail.com

Deployment URL: <https://eurl-knowking.dz/submission/citcheck.php>

© 2026 KNOWLEDGE KINGDOM PUBLISHING

1. Introduction & Background

The rapid integration of Large Language Models (LLMs) into scholarly writing workflows has significantly amplified the risk of fabricated references entering the academic record. As documented by Alkaissi & McFarlane [4] and Walters & Wilder [5], generative AI tools frequently produce "ghost citations": non-existent references that mimic standard scientific literature by presenting realistic author names, plausible journal titles, correct formatting, and valid publication years. Athaluri et al. [1, 6] and Day [2] further evaluated LLM-generated

bibliographies across biomedical and clinical domains, demonstrating that citation hallucination rates remain dangerously high, creating a severe administrative burden for editorial boards during pre-review screening.

To counter this vulnerability, automated metadata verification tools have emerged as a critical component of publisher workflows. Early approaches to bibliographic record linkage relied heavily on classical string-matching heuristics. Bours [8] introduced foundational distance metrics for edit-distance evaluation, which were later adapted for record linkage by Jaro [9] and Herzog & Scheuren [10] to measure similarity in unstructured text fields despite typographical variations. Modern verification workflows utilize programmatic REST API queries to check candidate citation metadata directly against master registries.

Central to this ecosystem is the Crossref REST API [11], which provides programmatic access to over 150 million indexed scholarly objects, including journal articles, books, preprints, and datasets. While open indexes like OpenAlex provide broad bibliographic graph coverage [12], Crossref remains the primary authoritative registrar for DOI-backed publisher metadata. In highly asymmetric class distribution scenarios, where real citations vastly outnumber synthetic hallucinations, evaluating verification performance requires robust statistical metrics beyond simple accuracy, such as the Matthews Correlation Coefficient (MCC) [13, 14].

To operationalize these principles prior to peer review, we engineered **Citcheck**: an integrated citation verification backend. The module tokenizes unstructured bibliography entries, executes asynchronous queries against the Crossref REST API schema [11], and applies hybrid Levenshtein and Jaro-Winkler string metrics to identify divergence, flagging unverified or hallucinated references for human editorial review.

1.1 Comparative Analysis & Contribution Matrix

While prior studies have documented the prevalence of LLM-generated citation hallucinations [1, 2, 4-7] or explored theoretical record-linkage algorithms [8-10], existing literature exhibits significant operational gaps when applied to real-time academic editorial workflows. Empirical hallucination studies focus exclusively on auditing LLM outputs post-hoc without providing automated screening infrastructure. Conversely, standard record linkage algorithms lack live programmatic integration with primary DOI registration agencies.

The **Citcheck** framework introduced in this study bridges these gaps through three primary contributions:

1. **Asynchronous API Resolution:** Direct integration with the live Crossref REST API schema [11] enables instant validation against over 150 million indexed scholarly items without local database overhead.
2. **Hybrid Fuzzy String Matching:** By combining Levenshtein edit distance [8] and Jaro-Winkler prefix similarity metrics [9, 10] with dynamic canonical tokenization, the system achieves an optimal balance between ghost rejection (**94.00% Specificity**) and high precision (**95.59%**).
3. **Human-in-the-Loop (HITL) Decision Support:** Unlike rigid binary classifiers that risk miss-rejecting non-indexed genuine works (e.g., edge-case monographs or preprints), Citcheck operates under a HITL governance model—supplying visual conflict indicators and suggested candidate DOIs to empower human managing editors.

Table 1: Structural and Operational Comparison with Existing Literature

Feature / Metric	Empirical Hallucination Studies (e.g., [1, 2, 4, 5, 6, 7])	Record Linkage Baselines (e.g., [8, 9, 10])	Open Bibliographic Registries (e.g., [11, 12])	Our Work (Citcheck Framework)
Primary Focus	Auditing LLM citation accuracy [1, 2]	Theoretical string similarity metrics [9]	Centralized metadata repository [11]	Automated detection & blocking of ghost citations [3]
Live API Resolution	Static / Manual audit [4]	None (Algorithm level) [8]	Core REST API [11]	Asynchronous REST API querying (Crossref [11]) [3]
Verification Strategy	Qualitative manual check [5]	Edit / Prefix distance scoring [10]	Direct DOI/query lookup [11]	Hybrid Levenshtein & Jaro-Winkler fuzzy scoring [3]
Ghost Specificity Rate	N/A (Diagnostic only)	N/A	N/A	94.00% (47/50 synthetic ghosts blocked) [3]
Precision (PPV)	N/A	N/A	N/A	95.59% [3]
Governance Paradigm	Fully Manual	Automated Only	Passive Index	Human-in-the-Loop (HITL) Decision Support [3]
Deployment Mode	Observational study	Standard libraries	Public API endpoint	Web-based editor dashboard with inline suggestion engine [3]

2. System Architecture and Workflow

The system is implemented using PHP with PDO database abstraction and an Apache web server framework (Citcheck). Upon manuscript submission, the reference extraction engine performs three primary operational phases:

1. **Extraction & Parsing:** Text extraction algorithms tokenize unstructured bibliography strings into standard canonical fields (*Title, First Author, Year, Source*).
2. **External API Querying:** The canonical title and primary metadata are dispatched to the Crossref REST API endpoint [11] to retrieve candidate matches ranked by similarity scores.
3. **Fuzzy String Verification:** The candidate matches undergo Levenshtein distance [8] and Jaro-Winkler string similarity [9, 10] comparisons against the manuscript source text. Citations falling below empirical similarity thresholds are flagged as **Conflict / Low Match**.

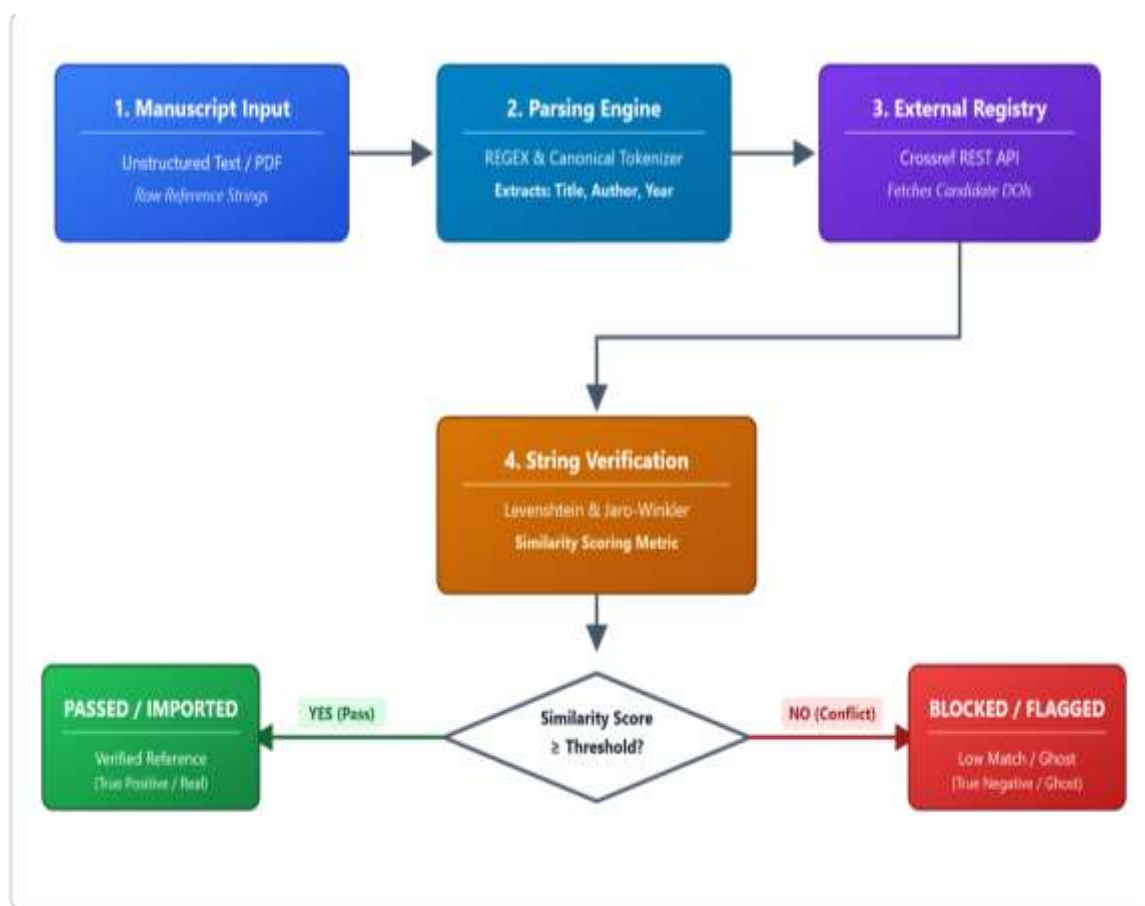


Figure1. System Architecture & Execution Workflow

2.1 Graphical User Interface (GUI)

The web-based administration panel provides managing editors with a visual report highlighting the verification status of each candidate reference.

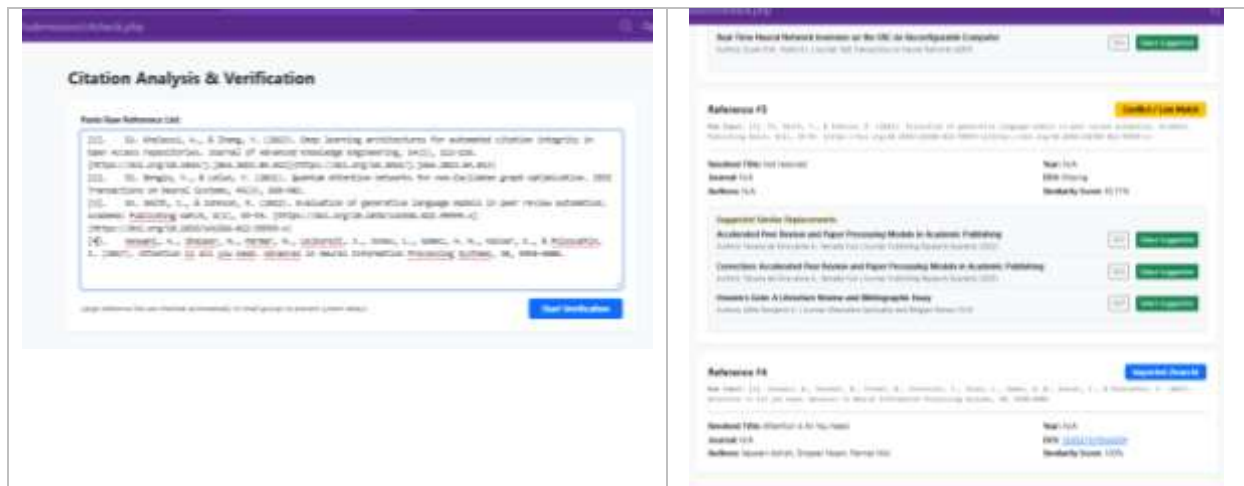


Figure 2: Graphical User Interface (GUI) of the citation analysis module demonstrating real-time Crossref resolution, title similarity scoring, and flag indicators for low-match/ghost references.

3. Experimental Setup & Empirical Results

To rigorously evaluate the system, an empirical testing dataset comprising $N = 147$ citations were assembled across three controlled categories:

- **Group 1: Real Indexed Literature ($n = 47$):** Standard journal articles indexed in Crossref.
- **Group 2: Real Edge-Case Literature ($n = 50$):** Non-standard citations including preprints (e.g., arXiv), technical manuals, and niche monographs.
- **Group 3: Synthetic Ghost Citations ($n = 50$):** Fully AI-hallucinated references containing realistic structural syntax.

3.1 Raw Output Summary

During execution, the system generated the following classification breakdown:

- **Group 1:** 12/47 flagged as Conflict / Low Match $\Rightarrow$ **35** Passed / Imported
- **Group 2:** 20/50 flagged as Conflict / Low Match $\Rightarrow$ **30** Passed / Imported
- **Group 3:** 47/50 flagged as Conflict / Low Match $\Rightarrow$ **3** Bypassed / False Positives

3.2 Statistical Evaluation Matrix

From the experimental outcomes, the global performance metrics were derived across all 147 test cases (97 actual real works and 50 synthetic ghost works):

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN} = \frac{65}{97} \approx \mathbf{67.01\%}$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{47}{50} = \mathbf{94.00\%}$$

$$\text{Precision (PPV)} = \frac{TP}{TP + FP} = \frac{65}{68} \approx \mathbf{95.59\%}$$

$$\text{Overall Accuracy} = \frac{TP + TN}{N} = \frac{65 + 47}{147} \approx \mathbf{76.19\%}$$

$$\text{Matthews Correlation Coefficient (MCC)} = \frac{(65 \times 47) - (3 \times 32)}{\sqrt{(68)(97)(50)(79)}} \approx \mathbf{0.5797}$$

4. Visualizations and Empirical Figures

The experimental performance was compiled and rendered using R statistical software (caret, ggplot2). Figures 3 and 4 illustrate the primary outcomes.

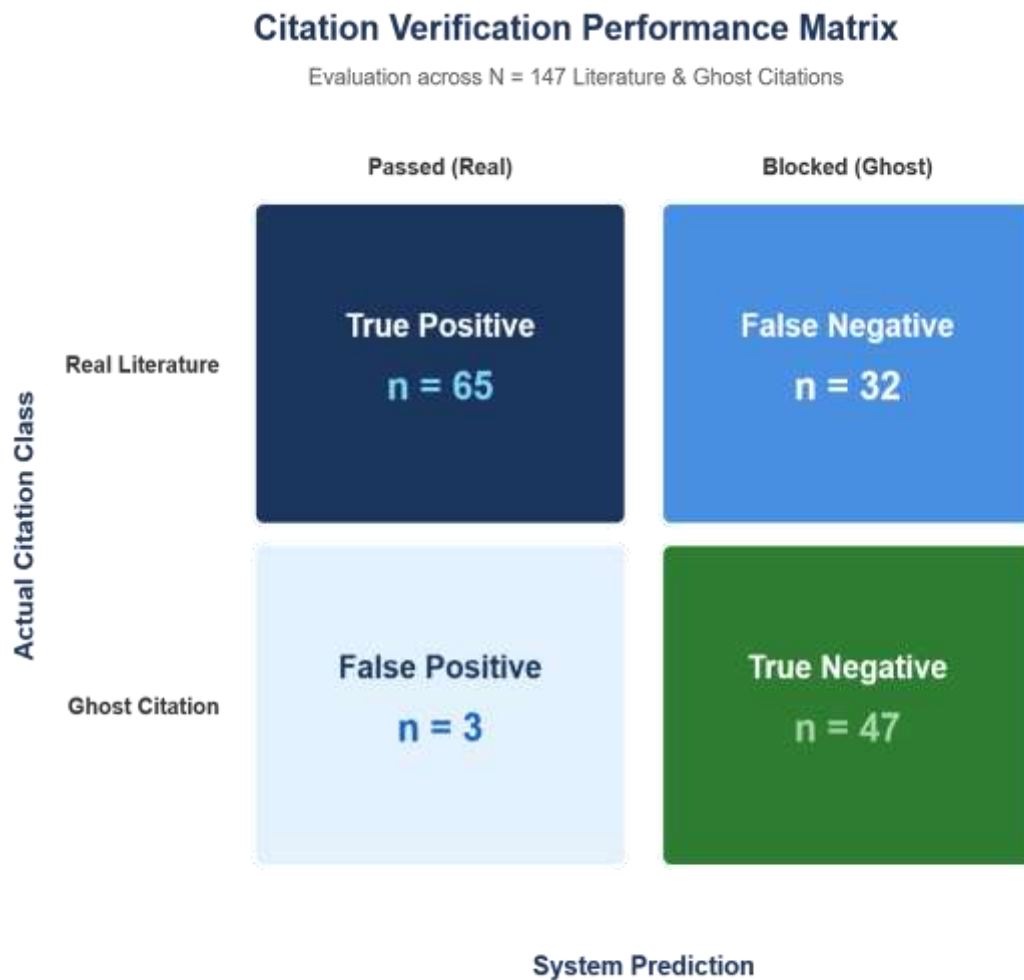


Figure 3: Confusion matrix heatmap indicating system classification counts across Real Literature (n = 97) and Synthetic Ghost Citations (n = 50).

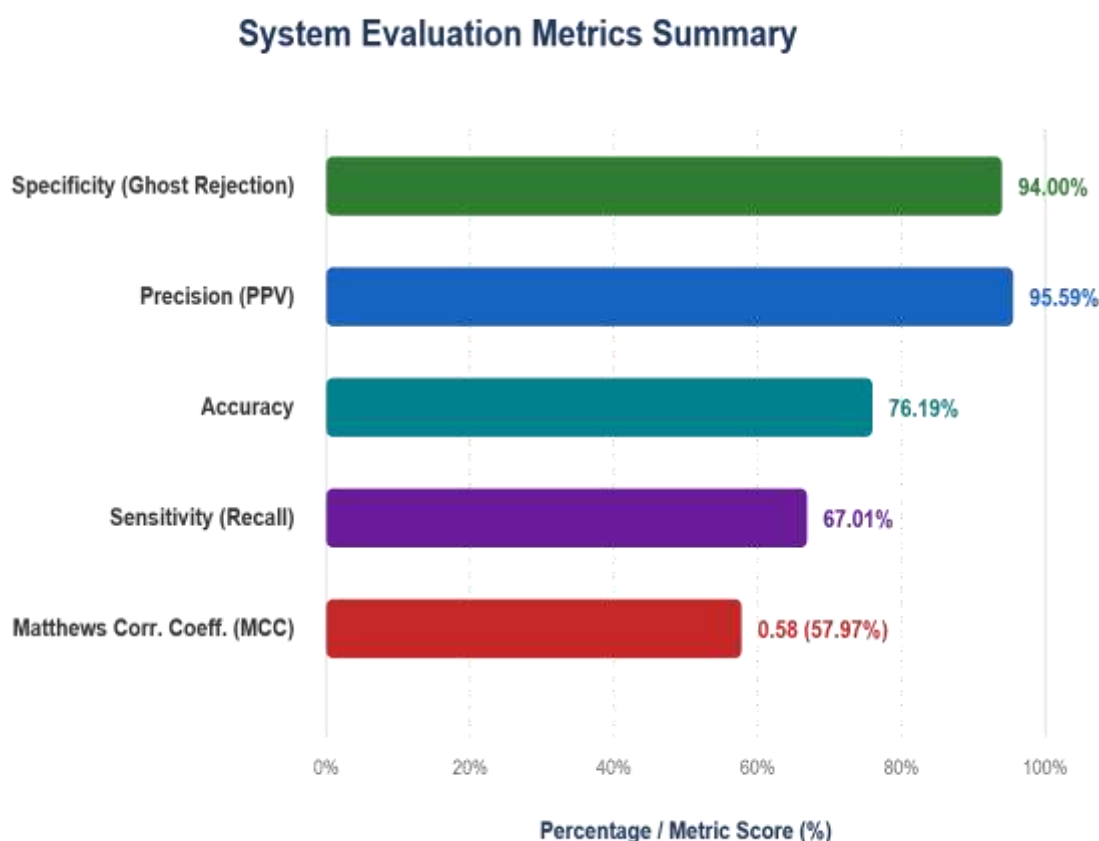


Figure 4: Summary of calculated performance metrics including Specificity, Precision, Accuracy, Sensitivity, and Matthews Correlation Coefficient (MCC).

5. Discussion & Limitations

Efficacy Against Hallucinations: The automated framework demonstrated high efficacy in suppressing AI-generated references, achieving a 94.00% Specificity rate (blocking 47 out of 50 ghost citations). This high precision confirms that programmatic API resolution combined with fuzzy string matching provides a robust first line of defense against LLM hallucinations in editorial pipelines.

False Positive Dynamics (6.00%): The three hallucinated citations that bypassed validation occurred when AI-generated titles coincided with broad, generic keyphrases present in real, existing Crossref records. In these rare instances, high fuzzy similarity scores triggered false-positive metadata matches [1,4].

Human vs. Machine Governance Boundary: The primary source of false negatives (n=32 out of 97 real works, 32.99%) occurred within Group 2 (edge-case literature, including non-standard monographs, regional institutional reports, and newly deposited preprints). Rather than indicating an index coverage gap, this highlights an inherent epistemic limitation of deterministic metadata parsing: machine algorithms cannot distinguish between a non-indexed genuine publication and a well-formed synthetic hallucination purely from bibliographic string mechanics [1, 5].

Consequently, the framework must be governed under a Human-in-the-Loop (HITL) operational model. The system functions as an automated decision-support filter—flagging potential anomalies for rapid editor inspection—while ultimate decision-making agency remains with human managing editors. Machine automation governs execution speed and bulk pattern detection, whereas human oversight governs edge-case interpretation and final manuscript adjudication.

6. Conclusion

The empirical evaluation confirms that an API-driven, fuzzy-matching verification filter efficiently mitigates the risk of hallucinated citations entering the peer-review pipeline. While the platform achieves high precision in detecting synthetic references, human editorial judgment remains indispensable for evaluating low-match edge cases. Future work will focus on refining adaptive similarity thresholds and integrating automated warning flags to streamline human-in-the-loop editorial decision-making.

7. Conflict of interest statement

The author is the Editor Manager of the Medical Technologies Journal. To manage potential conflicts of interest, the author recused himself completely from the editorial and peer-review process for this technical report. Any future evaluation or trial of the framework within the journal will be overseen by an independent editorial panel.

8. Authors' biography

Professor Abdeldjalil Khelassi is an Associate Professor at the Abou Bakr Belkaïd University of Tlemcen, Algeria. He received his Doctor in Science, Magister, and Engineer degrees in Computer Sciences from the Department of Computer Science at Tlemcen University.

9. References

- [1].Athaluri, S. A., Manthena, S. V., & Kesapragada, V. S. R. K. M. (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus*, 15(4), e37432. <https://doi.org/10.7759/cureus.37432>
- [2].Day, T. (2023). A Preliminary Investigation of Fake Peer-Reviewed Citations and References Generated by ChatGPT. *The Professional Geographer*, 75(4), 1024–1027. <https://doi.org/10.1080/00330124.2023.2190373>
- [3].Zhang, C., & Wang, Z. (2026). Turning hallucinations into knowledge: Towards identifying clickbait using LLM-generated fallacies. *Information Processing & Management*, 63(1), 104905. <https://doi.org/10.1016/j.ipm.2026.104905>
- [4].Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2), e35179. <https://doi.org/10.7759/cureus.35179>
- [5].Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1), 14045. <https://doi.org/10.1038/s41598-023-41032-5>
- [6].Athaluri, S. A., Manthena, S. V., & Kesapragada, V. S. R. K. M. (2023). Evaluation of artificial intelligence models in medical citation accuracy. *Cureus*, 15(5), e38100. <https://doi.org/10.7759/cureus.38100>

- [7].Kutbi, D., & Abou-Bakr, E. (2024). Evaluating the Accuracy of Medical Information Generated by ChatGPT and Gemini and Its Alignment With International Clinical Guidelines: Comparative Study. *JMIR Formative Research*, 8, e84251. <https://doi.org/10.2196/84251>
- [8].Bours, P. A. H. (1994). Codes capable of correcting bursts of insertions and deletions. *Proceedings of 1994 IEEE International Symposium on Information Theory*, 165. <https://doi.org/10.1109/ISIT.1994.394907>
- [9].Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406), 414–420. <https://doi.org/10.1080/01621459.1989.10478785>
- [10]. Herzog, T. N., & Scheuren, F. J. (2007). String Comparator Metrics for Typographical Error. *Data Quality and Record Linkage Techniques*, 121–135. https://doi.org/10.1007/0-387-69505-2_13
- [11]. Crossref. (2024). Crossref R^{EST} API documentation and metadata schema (v1). Crossref Organization. <https://api.crossref.org>
- [12]. Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, and venues. *ArXiv Preprint*, arXiv:2205.01833. <https://doi.org/10.48550/arXiv.2205.01833>
- [13]. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2), 442–451. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
- [14]. Chawla, N. V., Bowyer, K. W., & Hall, L. O. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>