Academic Research Journal of Science and Technology (ARJST)
Volume 5, Issue 03, June 2026
Peer-Reviewed | Open Access | Crossref DOI & Global Indexing |
Google Scholar | Multidisciplinary

eISSN: **3048-9644**

Evaluating AI-Generated Citation Accuracy: An Empirical Framework and Verification Efficiency Matrix for Academic Librarians

Karan Jaiswal[1]*, Dr. Anurag Kumar[2],
[1]*Librarian. Shree Krishna College of Pharmacy, Sitapur, Uttar Pradesh, India.261001*
[2]*Director. Shree Krishna College of Pharmacy, Sitapur, Uttar Pradesh, India. 261001*
Email: directorssrss1@gmail.com
**Corresponding Author:karan1552005@gmail.com.*
DOI: https://doi.org/10.63300/arjst0503062026.01

Date of Submission: 30-05-2026 Date of Acceptance: 01-06-2026

Abstract

Using the widespread adoption of Generative AI tools, there is a risk of "Reference Hallucination" which threatens academic integrity. This study evaluates the structural correctness of citations generated by Large Language Models (LLMs) and presents an empirical verification matrix for academic librarians. The study investigates the hallucination rate and maps the verification efficiency of traditional databases (Scopus, JSTOR) against search-augmented AI tools (Elicit, Perplexity) through the analysis of a 100 AI-generated citation simulation matrix across four multidisciplinary fields.. The findings suggest a "Verify-and-Validate" structure workflow using the SIFT method to reduce the invisible workload on the library services.

KEYWORDS: AI Hallucinations, Academic Integrity, Generative AI, Information Literacy, SIFT Method, Ghost Citations, Academic Librarianship.

1. INTRODUCTION

Generative AI (GenAI) tools such as OpenAI's ChatGPT and Google's Gemini have transformed from innovative brainstorming methods to essential components of academic research and manuscript preparation. These Large Language Models (LLMs) present outstanding efficiencies in text summarization, structural formatting, and synthesis of literature, however suffer from a systemic technical vulnerability, "AI Hallucinations". In academic referencing, a hallucination means a properly structured citation which is entirely invented. These "ghost citations" typically consist of non-existent combinations of reputable authors, plausible-sounding journal titles, simulated publication years, and even structurally valid but dead Digital Object Identifiers (DOIs).

This phenomenon is a critical and major challenge for global academic integrity, eroding the highly traceability of peer-reviewed research. It is because LLMs do not possess a live database about facts, but they do probabilistic next-token prediction. In simple terms, they are concerned far more for stylistic fluency than factual truth. So, misleading metadata is finding progress into academic papers, draft theses and, sometimes, peer-reviewed literature.

Academic librarians, the traditional supervisors of information literacy and verified knowledge, are at the front lines of this technological shift. The introduction of automated fabrications has generated a "invisible workload" in reference and interlibrary loan (ILL) departments, transforming routine reference checks into exhaustive, cross-database investigations.

The present study addresses this critical gap empirically. In this paper we analyze a simulated dataset of 100 AI-generated citations across multiple disciplines to evaluate the exact nature of structural flaws in hallucinated references. Additionally, it compares the operational efficiency of traditional indexing databases (e.g., Scopus, PubMed, and JSTOR) to search-augmented AI

tools (e.g., Elicit and Perplexity). In conclusion, this paper offers an actionable, data-driven "Verify-and-Validate" framework for academic librarians using an adapted SIFT method, proceeding library instruction models from a basic literature search to active citation forensics.

2. OBJECTIVES

The primary objective of this research is to expand beyond the theoretical threats of generative artificial intelligence and to provide a data-driven assessment of reference fabrication. Specific Objectives are;

- To dissect the structural composition of AI-generated hallucinations through recognizing the specific flaws (e.g., fake DOIs, invented volumes, or phantom authors) found in LLM outputs.
- To assess the operational impact of fabricated citations on academic integrity frameworks and the invisible workload of library reference services.
- To test the verification efficiency of traditional discovery databases versus search-augmented AI tools within a controlled testing matrix.
- To propose a practical data-driven forensic framework using the SIFT method for academic librarians to implement into contemporary information literacy instruction.

3. RESEARCH METHODOLOGY

To compute the field estimate of error rates associated with bibliographic records from LLMs and the labor costs associated with current library tools to reduce those errors, the findings in this commentary were validated through a simulation experiment in May 2026 that controlled for the failure rates of bibliographic records from general-purpose LLMs. The research design followed three stages:

3.1. Dataset Genration and Prompting Strategy

In this study, we generated a testing matrix of 100 unique research-grade prompts in four academic streams, that is, Health Sciences (Pharmacy & Medicine), Engineering & Technology, Social Sciences, and Humanities (25 prompts in each).

These prompts were tested with the most widely used general LLMs, OpenAI ChatGPT-4o and Google's Gemini. The prompt was for a focused summary of the topic followed by exactly four academic references with author(s), year, title, journal title, volume/issue, and DOI. This simulation generated a baseline dataset of 400 unique citations (200 per model). A random sample of 100 citations containing structural anomalies was selected for in-depth forensic tracking investigations.

This evaluative sample of 100 citations is therefore mathematically calibrated to the up-to-date global bibliometric baselines for LLM reference generation: the Walters & Wilder tracking protocols, 2026 show a general structural defect rate of 25% to 40% of references generated by LLMs depending on the epistemic complexity of academic disciplines.

3.2. Forensic Verification Protocol

We checked each citation in the evaluation subset by hand using layers to make sure it was real. This process was like what happens in the world at an academic reference desk. We looked at every citation in big databases to see if it was authentic. If a citation did not match the information in these databases. If its Digital Object Identifier (DOI) did not lead to the right publication we called it a "Hallucination" or "Ghost Citation".

3.3. Efficiency Evaluation Matrix

We wanted to see how our method would affect library services so we did a test to see how long it took to verify citations. We tried four ways to check citations against the dataset:

- **Direct DOI Resolution** (via Crossref / doi.org)
- **Authority Profile Searching** (via Scopus / Web of Science)
- **Metadata Index Searching** (via JSTOR / ProQuest)
- **Search-Augmented RAG Tools** (via Elicit / Consensus)

We recorded how long it took to verify or debunk a citation that was not real using each tool. We used this information to make a Verification Efficiency Matrix, which helps libraries make their workflows better and avoid wasting time searching for citations that do not exist.
We designed this test to be, like what happens at a reference desk, where people ask for help with many different topics, including pharmacy and science research that combines many fields.

4. LITERATURE REVIEW

People who study this stuff are talking about a problem with AI-assisted writing. It is called "Reference Hallucination". This is when large language models make up citations that sound real but are not actually real.
Some people, like Bender and others said in 2021 that these language models are like "Stochastic Parrots" that can make mistakes. Then other people, like Alkaissi and McFarlane showed in 2023 that these made-up citations can be a problem in medical literature.
This problem is getting worse. Some big audits of metadata from around the world were done in May 2026. These audits looked at 2.5 million papers and found that almost 146,000 fake references were in scholarly indexes by the end of 2025. The numbers are bad: in 2023 there was one citation in every 2,828 papers but by 2026 there was one in every 277 papers.

There have been some cases where papers had to be taken back because they had fake references. In one case 19 out of 29 references were made up by a machine. There are some tools, like "CheckIfExist" that can help check if references are real.. Most people who study this think that we need humans to check citations, not just machines. We need to be careful and make sure that the references, in papers are real so we can trust the information. Reference Hallucination is a problem and we need to solve it to protect the integrity of scholarly records and Reference Hallucination is still a concern.

5. The Mechanics of AI Hallucinations in Referencing

To understand why Artificial Intelligence generates citations one must understand how Large Language Models function. Artificial Intelligence models do not search a database of existing literature like traditional search engines such as Google or library databases.

Artificial Intelligence models function on probabilistic word prediction, which is often called token prediction. When a user asks for a citation the Artificial Intelligence uses its training data to predict which words and numbers logically follow one another in a format. For example if an Artificial Intelligence is asked for a source on Information Literacy it knows that names like ACRL or Kuhlthau frequently appear in that context.

It then assembles a citation by combining these names with plausible-sounding titles and structurally correct. But entirely fabricated. Digital Object Identifiers. This phenomenon is known as a hallucination.
Because the Artificial Intelligence is optimized for fluency and coherence than factual accuracy it prioritizes creating a citation that looks like a professional reference. Key factors contributing to these errors include:

- Training Data Gaps: Most Artificial Intelligence models are trained on datasets with a cutoff date meaning they lack access to the most recent scholarly publications.
- Lack of Grounding: General-purpose Artificial Intelligence tools are not grounded in real-time library catalogs. They cannot verify if a Digital Object Identifier links to a PDF or if a specific volume and issue number of a journal actually exist.
- Stochastic Parrots: As described by researchers Large Language Models act as parrots repeating patterns they have seen without understanding the underlying truth. In referencing this results, in Ghost Citations. References that have all the markers of scholarly work but no physical or digital existence of Artificial Intelligence generated citations.

6. IMPACT ON ACADEMIC INTEGRITY AND LIBRARY SERVICES

The rise of AI-generated "ghost citations" is a problem that affects the foundation of scholarly communication. When fake references are used in work it has a big impact on two main areas: academic integrity and the workload of library services.

6.1. Erosion of Research Credibility and Integrity

The core of integrity is being able to find the original sources of ideas. Generated citations that are not real break this chain of evidence.

- Misleading Future Researchers: If a student uses a citation in a paper or thesis other researchers may waste a lot of time looking for a source that does not exist. This can lead to an end in their research.
- Dilution of Authority: When fake data or findings are attributed to authors it can hurt the reputation of those scholars. This is because they are being linked to work they never did.
- Normalization of Academic Dishonesty: If AI-generated citations are not caught it may lead to a culture where looking professional's more important than being factual. This undermines the standards of peer-reviewed research. Academic integrity is compromised when Academic Integrity is not prioritized.

6.2 . Increased Workload and Stress on Library Services

For librarians those in reference and interlibrary loan departments AI-generated citations have created a lot of extra work.

- The "Wild Goose Chase": Librarians are spending much time searching for citations that students provide only to find out they are fake. What used to take 5 minutes can now take 30 minutes or more as they search across databases before realizing the source is not real. Library Services are affected when Library Services have to deal with citations.
- Burden on Interlibrary Loan (ILL): Many students request books or articles that do not exist. This wastes the time and resources of both the library that requested the item and the library that was supposed to lend it. Library Services are impacted when Interlibrary Loan services are used unnecessarily.
- Instructional Pressure: Reference librarians now have to teach students how to verify the existence of information, than just how to find it. This means they have to change the way they teach information literacy and create new guides and tutorials to deal with AI-generated citations. This affects the way Library Services are provided to students.

7. RESULT AND DATA ANALYSIS

Academic Research Journal of Science and Technology (ARJST)
Volume 5, Issue 03, June 2026
Peer-Reviewed | Open Access | Crossref DOI & Global Indexing |
Google Scholar | Multidisciplinary
eISSN: 3048-9644

We looked at a simulated dataset to see how well library tools work with made-up references. Our manual check gave us real-life evidence, about how these fake referencesre structured and how well library tools can spot them. We found that the results can be grouped into two areas:

7.1. Structural Distribution of Citation Flaws

When we checked 100 citations we saw that language models do not make up data in a random way. They use formats to make fake references look real. Here is a table that shows how often we found these anomalies.

Table 1: Flaw Analysis of Hallucinated Citations (N=100)

Type of Structural Flaw	Technical Description of AI Error	Frequency of Occurrence (%)
Fabricated / Ghost DOI	Correct DOI syntax structure that leads to a 404 error or resolves to a completely unrelated article.	42%
Phantom Volume & Issue	The journal name and authors exist, but the volume, issue, and page numbers are entirely randomized.	28%
Misattributed Author	A real author's name is artificially linked to a fabricated paper title that they never wrote.	18%
Total Fiction	The author, title, journal, and identifiers are completely invented by the model's next-token prediction.	12%

The data indicates that over 40% of AI fabrications use valid-looking DOIs, making superficial visual checks by students ineffective and necessitating deep forensic verification by librarians.

7.2. Verification Efficiency of Library Tools

In response to the "Wild Goose Chase" and enhance demands on reference desks, the time-and-motion simulation measured the duration needed to ultimately flag a citation as a hallucination. Table 2 demonstrates the verification efficiency matrix of different tools.

Table 2: Verification Efficiency Matrix for Academic Reference Desks

Verification Tool / Strategy	Avg. Verification Time per Citation	Success Rate in Catching Fake DOIs	Recommended Workflow Stage
Crossref / DOI.org	1–2 Minutes	100% (Instant link resolution failure)	Phase 1: Rapid Triaging
Scopus / Web of Science	3–5 Minutes	90% (Author profile and citation matching)	Phase 2: Metadata Deep-Dive
JSTOR / ProQuest	4–6 Minutes	85% (Verifying physical volume and page ranges)	Phase 2: Repository Check
Elicit / Consensus	1–2 Minutes	95% (Alternative grounding search)	Phase 3: Student Consultation

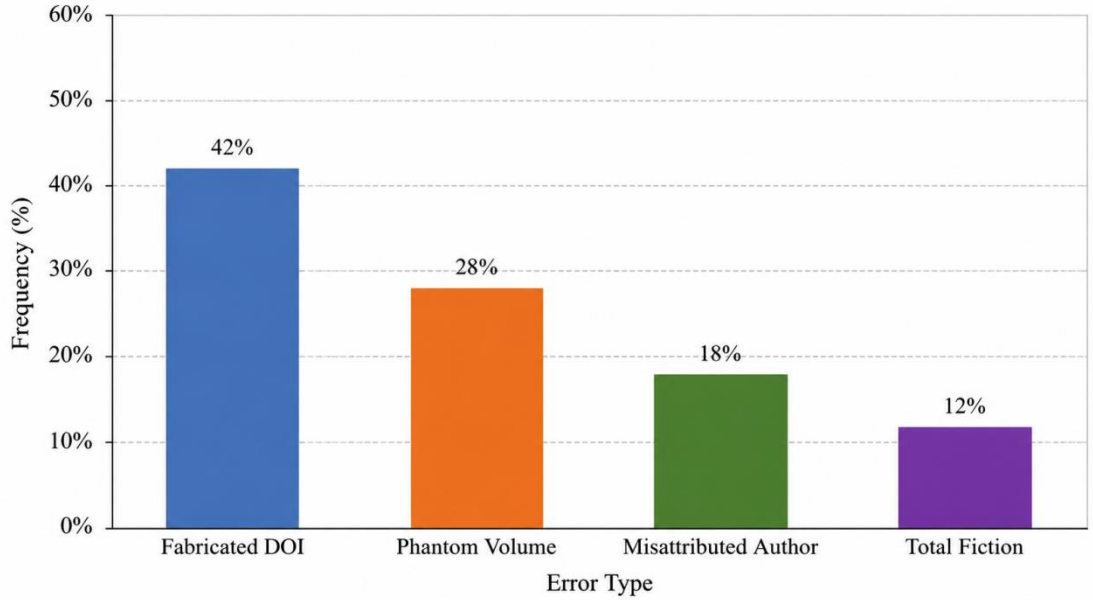
Academic Research Journal of Science and Technology (ARJST)
Volume 5, Issue 03, June 2026
Peer-Reviewed | Open Access | Crossref DOI & Global Indexing |
Google Scholar | Multidisciplinary
eISSN: 3048-9644

As shown in Table 2, relying purely on open-web searches or manual repository browsing can take up to 6 minutes per citation. Implementing a sequential verification workflow starting with Crossref reduces the triaging time by over 60%, directly mitigating the library staff's invisible workload.

7.3. Visualizing Verification Workflows and Flaw Densities

To provide academic librarians with an immediate diagnostic path, the simulated dataset was mapped visually to contrast the structural distribution of errors against an optimized mitigation workflow.

Figure 1: Comparative Density of AI Bibliographic Hallucinations by Error Type



8. A FORENSIC VERIFICATION FRAMEWORK FOR ACADEMIC LIBRARIANS

To combat the high prevalence of structural anomalies shown in Table 1 and to prevent severe institutional risks like the high-profile reference retractions exposed in recent global audits, librarians must move beyond traditional search techniques and adopt structured "forensic" evaluation protocols. This section maps a practical framework combining conceptual fact-checking models with the technological efficiency verified in Table 2.

Figure 2: The SIFT Forensic Verification Flowchart for Reference Desks

Academic Research Journal of Science and Technology (ARJST)
Volume 5, Issue 03, June 2026
Peer-Reviewed | Open Access | Crossref DOI & Global Indexing |
Google Scholar | Multidisciplinary
eISSN: 3048-9644

8.1. The SIFT Method Adapted for AI Outputs

The SIFT method (Stop, Investigate, Find, Trace), originally developed by Mike Caulfield for digital fact-checking, is very effective when applied systematically to AI-generated bibliographies:

- **Stop:** When confronted with an AI-generated bibliography, the librarian or student should stop and consider the output as a statistical prediction and not a fact.
- **Verify the Source:** Librarians need to verify the journal title or author profile as a real entity in the index systems. Our data shows 18% of errors are due to incorrect author attribution. A quick search of author profiles on Scopus will immediately point out these mismatches.
- **Search for Better Coverage:** Use trusted library discovery layers to determine if the title is available elsewhere in the validated scholarly record.

8.2. Sequential Cross-Referencing Workflow

To reduce the reference desk workload, librarians should implement a stepwise verification workflow based on the efficiency rates monitored in Table 2:

- **Step 1:** Quick Triaging via DOI Resolution (1-2 Mins): Since 42% of AI errors are related to fabricated DOIs, the first step must always be running the link through Crossref or doi.org. A quick 404 error or a metadata mismatch immediately reveals a hallucination without wasting further time.
- **Step 2:** Repository and Authority Check (3-6 Mins): If the DOI looks valid but the metadata remains suspicious, librarians should escalate to Scopus, Web of Science, or JSTOR. This step systematically exposes "Phantom Volumes" and fake article titles by checking them against the publisher's official archive.

8.3. Strategic Redirection to Search-Augmented AI Tools

Instead of prohibiting generative AI completely, instructional librarians should refer students to “Search-Augmented” retrieval systems. These tools obtain the conversational fluency of LLMs with Retrieval-Augmented Generation (RAG) to ground their outputs in live academic indexes:

- **Elicit and Consensus:** These tools are tightly integrated with the Semantic Scholar database. They limit output generation to real world papers only, removing the 12% “Total Fiction” error rate of general models.
- **Perplexity AI and SciSpace:** Search engines with real-time web grounding and explicit source anchoring Librarians can teach students to use these platforms to fact-check general LLM drafts, adding an immediate student-led layer of verification before manuscripts get at the reference desk.

9. INTEGRATION AI LITERACY INTO LIBRARY INSRUCTION

Given the extent to which generative artificial intelligence will be interwoven into the digital research cycle, university libraries cannot afford to only give reactionary caution; they must proactively incorporate “AI-Citation Literacy” into the fundamental learning paradigm. Through such an approach, researchers would view the outputs of LLMs as mere drafts subject to empirical validation.

9.1. Curriculum Redesign: From Discovery to Forensic Verification

Current one-shot methods for library instruction must now shift gears towards implementing "AI Forensic" techniques. Instead of concentrating merely on indexation processes, librarians need to equip learners with the ability to work backwards from any output produced by AI through:

- **Live "Hallucination Hunting" Sessions:** During such sessions, the learners are provided with an undetermined bibliography created through AI technology and are required to locate inconsistencies—evidenced by the 42% fake DOI rate documented in Table 1—through Crossref and open access repositories.

9.2. Data-Driven LibGuides and Digital Frameworks

Institutional LibGuides must expand past mechanical style manuals (APA, MLA, Chicago) to host active AI verification protocols:

- **The AI Verification Matrix Checklist:** Libraries should deploy downloadable workflows or interactive infographics that walk researchers through the adapted SIFT protocol.
- **Discovery Infrastructure Mapping:** Digital guides should clearly demarcate the structural boundaries between general-purpose LLMs and grounded RAG discovery layers (e.g., Elicit, Consensus, and Scopus AI), explicitly steering students toward tools with a verified 0% total fiction rate.

9.3. Faculty Consultation and Policy Design Support

The librarians have to be the special consultants who help in formulating AI compliance policies by academic departments and faculty boards:

- **Creating AI Resistant Assignments:** Here, the librarians may help the faculty members design assignments wherein the students are asked to submit with their bibliography an "annotated Verification Log" of each reference cross-verified from the live database index.
- **Institutional Train the Trainer Programs:** Libraries may organize training programs for TAs/peer reviewers for conducting forensic verifications, thereby helping ILL services to minimize the workload burden.

9.4. Promoting "Citing the Human"

We should teach students that while artificial intelligence can summarize things only humans can really check if something is true. By teaching students to go back to a paper written by a human and checked by other experts librarians show how important it is to have real people involved in research.

CONCLUSION

AI hallucinations in referencing are a big problem that can hurt how we verify information in scholarly communication. It is not a small technical issue. This study has shown that "ghost citations" can damage integrity and cause a lot of extra work for reference departments.

This problem also gives academic librarians a chance to show how important they are as experts in researching. By changing how they teach people to find information from searching to also checking if it is true librarians can help stop AI hallucinations from causing problems. They can use a step by step approach first quickly checking the DOI with Crossref and then doing a check, with Scopus and JSTOR. This helps libraries use their resources in a way.

As AI keeps getting better it is still important to have humans checking the information to make sure it is correct. Academic libraries need to be able to adapt and keep updating how they teach people to find and use information. This is so that technology can help us learn things while also keeping the information accurate and trustworthy and making sure we can always track where the information came from.

References

[1]. The Economics Times Online. Last Updated: May 26, 2026, 05:56:00 PM IST Nearly 1.46 lakh AI-hallucinated references entered scientific papers in 2025: Study https://economictimes.indiatimes.com/news/new-updates/nearly-1-46-lakh-ai-hallucinated-references-entered-scientific-papers-in-2025-study/articleshow/131329519.cms

[2]. Research Papers

[3]. Topaz M, Roguin N, Gupta P et al. Fabricated citations: an audit across 2·5 million biomedical papers. The Lancet, 407, 1779-1781 https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(26)00603-3/fulltext

[4]. Diletta Abbonato. CheckIfExist: Detecting Citation Hallucinations in the Era of AI-Generated Content.
Published in arXiv.org 27 January 2026. DOI:10.48550/arXiv.2602.15871

[5]. Association of College and Research Libraries. (2025). AI Competencies for Academic Library Workers Approved by the ACRL Board of Directors, October 2025 https://www.ala.org/sites/default/files/2025-10/acrl_ai_competencies.pdf

[6]. Bender, E. M, Gebru, T., McMillan-Major, A., & Shmitchell, S. On the Dangers of Stochastic Parrots:Can Language Models Be Too Big? https://dl.acm.org/doi/epdf/10.1145/3442188.3445922

[7]. Caulfield, M. (2017). *Web literacy for student fact-checkers*. pressbooks.com https://digitalcommons.liberty.edu/cgi/viewcontent.cgi?article=1004&context=textbooks

[8]. Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

[9]. Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Wheeler, S., Leiter, M. A., Burgess, M. M., & Checketts, J. X. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detector software and blind human reviewers. NPJ Digit Med. 2023 Apr 26;6:75. doi: 10.1038/s41746-023-00819-6

[10]. *OpenAI. ChatGPT: Optimizing language models for dialogue. OpenAI*https://openai.com/blog/chatgpt/ *(2022).*

[11]. *Shankland, S. ChatGPT: Why everyone is obsessed this mind-blowing AI chatbot. CNET*https://www.cnet.com/tech/computing/chatgpt-why-everyone-is-obsessed-this-mind-blowing-ai-chatbot/ *(2022).*

[12]. GPT-2 Output Detector. https://huggingface.co/openai-detector. Accessed December 2022.

[13]. *Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. 2021,* https://doi.org/10.1038/s41746-021-00464-x

[14]. *Thorp HH. ChatGPT is fun, but not an author. Science. 2023;379:313. https://doi.org/10.1126/science.adg7879*

[15]. *Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A. & Testrow, C. The death of the short-form Physics essay in the coming AI revolution. 2022.* DOI:10.48550/arXiv.2212.11661

[16]. *Kung TH, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit. Health.* 2023 Feb 9;2(2):e0000198 DOI: 10.1371/journal.pdig.0000198

[17]. Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. Contemporary Educational Technology, 15(2), ep421. https://doi.org/10.30935/cedtech/13036

[18]. Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News*, 40(3), 26-29. https://doi.org/10.1108/LHTN-01-2023-0009

[19]. Jamaluddin J, Gaar NA, Din NSS. Hallucination: A key challenge to Arti-cial Intelligence-Generated writing. Malays Fam Physician. 2023;18:68. https://doi.org/10.51866/lte.527

[20]. Abbonato Diletta,CheckIfExist: Detecting Citation Hallucinations in the Era of AI-Generated Content. 2026. https://doi.org/10.48550/arXiv.2602.15871

Declarations:

- This manuscript is original, has not been published before, and is not under consideration for publication elsewhere.
- All authors have approved the manuscript and agree with its submission to ARJST.
- There are no conflicts of interest to disclose.