BG Pulse Journal USA
ISSN:1947-2536 http://www.bgpulseusa.com/index.php/bg

DETECTING FACTUAL ERRORS AND FABRICATED SCIENTIFIC REFERENCES IN GENERATIVE AI RESPONSES: THEORETICAL FOUNDATIONS AND RESEARCH PROSPECTS AMONG UNIVERSITY STUDENTS IN UZBEKISTAN

Rakhimkhonova Diyora M.

Sayfullayeva Shirin B.

Student, Namangan Branch of KIUT, Namangan, Uzbekistan

rakhimkhanovadiyora@gmail.com

dr.shirinbaxramovna@gmail.com

+998(99)-475-35-38

+998(93)-278-82-29

Doi: https://doi.org/10.5281/zenodo.23852946

Abstract. Generative artificial intelligence (AI) tools are increasingly used by university students to obtain explanations, summarise literature, prepare assignments and search for scientific sources. However, these systems can produce, in a fluent and convincing style, incorrect facts, inaccurate bibliographic details and references to publications that do not exist. Drawing on literature on AI hallucinations, information literacy, evidence-based practice and cognitive psychology, this article analyses the theoretical foundations of detecting factual errors and fabricated references in generative AI output. Empirical studies show that a substantial proportion of references generated by large language models are fabricated or inaccurate, and that even medical trainees do not reliably recognise such errors. The article proposes a conceptual framework for studying students' ability to distinguish accurate from inaccurate information and genuine from fabricated references, a multicentre study design for universities in Uzbekistan, and a six-step educational model for improving AI literacy.

Keywords: generative artificial intelligence, AI hallucinations, fabricated references, information literacy, critical thinking, higher education, medical education, academic integrity.

Introduction

Generative artificial intelligence has rapidly become part of everyday academic practice. Large language models (LLMs) such as ChatGPT are used by students to explain complex concepts, summarise texts, generate ideas and draft written work [12; 20]. In medical education, interest in these tools increased after ChatGPT reached or approached the passing threshold of all three steps of the United States Medical Licensing Examination without specialised training [14]. A systematic review of ChatGPT in healthcare education, research and practice identified potential benefits alongside serious concerns, including inaccurate content, fabricated citations and risks to academic integrity [20].

The key limitation of LLMs is their capacity to generate plausible but false information — a phenomenon commonly referred to as "hallucination" [10]. LLMs generate text by predicting likely

sequences of words on the basis of patterns learned from training data [19]; they do not, by default, verify statements against a reliable knowledge base. As a result, a fluent and confident answer may contain incorrect facts, unsupported conclusions, inaccurate descriptions of studies, or references to publications that do not exist [2; 4].

This problem is particularly serious in academic work. A student who receives a convincing explanation accompanied by a list of apparently relevant references may incorporate them into an assignment, presentation or thesis without checking them. In medicine, such errors can distort professional knowledge and, potentially, future clinical decisions [9].

Uzbekistan has made the development of artificial intelligence a national priority: the Strategy for the Development of Artificial Intelligence Technologies until 2030, approved in October 2024, envisages the wide introduction of AI in various sectors, including education [16]. This makes it important to determine whether university students are prepared to evaluate AI-generated information critically. To date, however, no published empirical study has examined this question among students in Uzbekistan.

The aim of this article is to analyse the theoretical foundations of detecting factual errors and fabricated references in generative AI responses and to propose a framework and study design for investigating this competence among university students in Uzbekistan.

The article addresses three research questions:

1. What types of factual errors and reference inaccuracies occur in generative AI responses, and how frequent are they?
2. Which information literacy and critical appraisal skills enable students to identify these problems?
3. How can students' ability to evaluate AI-generated information be measured in Uzbek higher education?

Materials and methods

The study uses a narrative review and conceptual analysis. Sources were identified in PubMed, Scopus and Google Scholar using combinations of the terms "ChatGPT", "large language model", "hallucination", "fabricated references", "citation accuracy", "information literacy", "AI literacy" and "medical education". Publications from 2019 to 2025 were considered for empirical studies; foundational works on information literacy, research methodology and cognitive psychology were included regardless of year.

Priority was given to (a) empirical studies that quantified the accuracy of references or facts generated by LLMs; (b) studies that assessed human ability to detect AI-generated errors; (c) frameworks and guidance on information literacy and responsible AI use in education. All references used in this article were verified against publisher records or DOI registration data.

The article does not report original survey or experimental data. The proposed instrument and study design are intended for a future empirical study.

Literature review and theoretical foundations

1. Types of unreliable AI-generated information

For academic purposes, unreliable AI-generated content can be divided into six categories (Table 1). These categories are not interchangeable: a reference may be real but cited for a claim it does not support, while a fabricated reference may contain plausible authors, a real journal title and a correctly formatted DOI.

Table 1. Types of unreliable information in generative AI responses

Type	Description	Example
Factual error	Incorrect date, definition, figure, anatomical or physiological description	Wrong origin or insertion of a muscle
Unsupported claim	Statement presented confidently without sufficient evidence	"It is proven that…" without a source
Misrepresentation of research	Inaccurate summary of study methods, results or conclusions	Reporting a pilot study as a randomised trial
Bibliographic inaccuracy	Errors in authors, title, year, journal, volume, pages or DOI	Real article with wrong year and DOI
Fabricated reference	Citation that does not correspond to any existing publication	Plausible title in a real journal that does not exist
Misattributed reference	Real publication cited as evidence for a claim it does not support	Review on another topic cited as proof

Source: developed by the author based on [2; 4; 10; 21].

Evaluating AI-generated academic material therefore requires two distinct activities: checking whether the **claims** are accurate, and checking whether the **cited sources** exist and actually support those claims.

2. Empirical evidence on the accuracy of AI-generated references

A growing number of studies have quantified the problem (Table 2). Walters and Wilder asked ChatGPT to write short literature reviews and found that 55% of the references produced by GPT-3.5 and 18% of those produced by GPT-4 were fabricated; many of the genuine references also contained substantive errors [21]. Bhattacharyya et al. analysed 115 references in ChatGPT-generated medical papers: 47% were fabricated, 46% were authentic but inaccurate, and only 7% were both authentic and accurate [4]. Chelli et al. compared LLMs in the context of systematic

reviews and reported hallucination rates of 39.6% for GPT-3.5, 28.6% for GPT-4 and 91.4% for Bard [6]. Gravel et al. found that ChatGPT's answers to medical questions frequently included fabricated references that appeared credible [8]. Case reports by Alkaissi and McFarlane and by Athaluri et al. also describe fabricated citations in AI-assisted scientific writing [2; 3].

Table 2. Selected empirical studies on reference accuracy and error detection

Study	Design and material	Main finding
Walters & Wilder, 2023 [21]	Literature reviews generated by GPT-3.5 and GPT-4	55% (GPT-3.5) and 18% (GPT-4) of references fabricated
Bhattacharyya et al., 2023 [4]	115 references in ChatGPT-generated medical papers	47% fabricated, 46% inaccurate, 7% accurate
Chelli et al., 2024 [6]	References generated for systematic reviews	Hallucination rate 39.6% (GPT-3.5), 28.6% (GPT-4), 91.4% (Bard)
Gravel et al., 2023 [8]	ChatGPT answers to medical questions	High proportion of fabricated but credible-looking references
Zhou et al., 2025 [22]	General practice trainees in China assessing AI answers	Mean accuracy of hallucination detection 55.0%
Pratama, 2025 [18]	Training session for students on verifying ChatGPT references	Targeted training as a feasible educational intervention

Source: compiled by the author.

Two conclusions follow. First, although newer models produce fewer fabricated references, the problem has not disappeared. Second, error rates depend on the model, the task and the field, so findings from one study cannot be applied directly to another context.

Evidence on **human detection** of AI errors is scarcer. Zhou et al. examined general practice trainees in China and found a mean accuracy of 55.0% in identifying hallucinations in AI-generated answers — only slightly better than chance for a binary judgement [22]. Pratama described a training session in which students learned to identify and correct fabricated references in ChatGPT-generated literature reviews, illustrating the feasibility of targeted instruction [18]. Neither study examined students in Central Asia.

3. Scientific reference verification

Scientific references allow readers to trace claims back to their sources. A reference should not be accepted merely because it contains a plausible title, author list, journal and DOI.

Verification should establish (a) that the publication exists and (b) that its content supports the associated claim. A practical procedure includes five steps:

1. search for the publication by its exact title in a scholarly database;
2. compare the authors, year, journal, volume, issue and pages with the publisher's record;
3. resolve the DOI through doi.org or Crossref and check that it leads to the same publication;
4. read the abstract or full text and assess whether it supports the claim;
5. check whether the study type is appropriate for the conclusion drawn.

For biomedical literature, PubMed is the primary resource; Crossref provides bibliographic metadata, and Google Scholar helps locate publications across disciplines. These services differ in coverage, and none of them alone guarantees validity.

4. Information literacy and critical appraisal

The Framework for Information Literacy for Higher Education describes information literacy as a set of integrated abilities for discovering, evaluating and using information responsibly, emphasising that "authority is constructed and contextual" [1]. An AI-generated answer should therefore be treated as information to be evaluated, not as an authoritative source in itself.

Critical appraisal is especially important in medicine and dentistry. Greenhalgh and Dijkstra stress the need to understand study design and to evaluate evidence before applying it in practice [9]. Students should learn to ask not only whether an AI answer sounds reasonable, but also whether it is supported by reliable evidence, whether the source is appropriate and whether contradictory findings exist.

5. Cognitive factors

Human judgement is shaped by fast intuitive processes and slower deliberate reasoning [11]. Fluent, well-structured text tends to be processed easily and may be accepted with less scrutiny. Research on misinformation suggests that susceptibility to false content is associated more with insufficient deliberation than with motivated reasoning [17]. In addition, people with limited competence in a domain often overestimate their own performance [13]. For AI literacy, this implies that students' self-reported confidence in detecting errors may not match their actual ability, and that educational strategies should explicitly prompt deliberate verification.

6. Responsible use of generative AI in education

UNESCO's guidance on generative AI in education and research advocates a human-centred approach and calls on institutions to develop policies, build capacity and validate AI tools before use [15]. Responsible use does not require prohibiting AI-assisted learning; rather, students must understand the limitations of these systems and remain accountable for the accuracy of their work [12]. Assessment methods that reward verification and reasoning — rather than the reproduction of generated text — support this goal.

Proposed research framework

1. Competencies to be assessed

A future multicentre study is proposed to assess two primary competencies:

- **factual error detection** — the ability to distinguish accurate scientific statements from inaccurate or unsupported ones;
- **reference verification** — the ability to identify genuine publications, detect fabricated or materially inaccurate references and determine whether a source supports a claim.

Secondary aims include examining associations between performance and prior AI use, academic discipline, year of study, information literacy training and familiarity with scholarly databases, as well as comparing self-rated confidence with actual accuracy.

2. Study design and sample

A cross-sectional assessment is proposed among students of several universities in Uzbekistan, including medical and non-medical institutions. A cross-sectional design is appropriate for describing the level of a competence and its correlates at one point in time, whereas causal questions — such as the effect of training — require an intervention design [5; 7]. Stratified cluster sampling (by university and year of study) is recommended. The sample size should be calculated on the basis of the primary outcome — the proportion of correctly classified items — with an assumed proportion of 0.5 (the most conservative assumption), a precision of ±5%, a 95% confidence level and a design effect for clustering. Under these assumptions, approximately 385 participants would be needed without clustering; the final figure should be adjusted for the design effect and the expected response rate.

3. Assessment instrument

A 24-item test with four blocks is proposed (Table 3). Items should be developed in the language of instruction (Uzbek, Russian or English) and checked independently by at least two subject experts.

Table 3. Structure of the proposed assessment instrument

Block	Items	Content
A. Accurate statements	6	Correct statements on the curriculum (e.g. anatomy, physiology), with sources
B. Statements with errors	6	Statements containing a factual error or unsupported claim
C. Genuine references	6	Real publications correctly cited for a relevant claim

Block	Items	Content
D. Fabricated or inaccurate references	6	Non-existent publications (3) and real publications with major errors or misattribution (3)

Source: developed by the author.

For each item, participants indicate whether the statement or reference is correct and rate their confidence on a five-point scale. Before the main study, the instrument should be piloted (approximately 30–50 students) to assess clarity, difficulty and discrimination; internal consistency (e.g. Kuder–Richardson 20) and item analysis should be reported.

4. Data analysis

The analysis should report overall accuracy, accuracy by block, sensitivity and specificity for detecting errors and fabricated references, item-level difficulty, and the relationship between confidence and accuracy. Associations with student characteristics can be examined using multilevel logistic or linear regression to account for clustering by university. Participation must be voluntary and anonymous, and ethical approval should be obtained from the relevant institutional committee.

Proposed educational model

Based on the reviewed literature, a six-step model for developing AI literacy is proposed (Table 4). The model can be integrated into courses on research methodology, academic writing, evidence-based medicine and discipline-specific teaching.

Table 4. Six-step educational model for evaluating AI-generated information

Step	Name	Student activity
1	Recognise limitations	Learn that fluent, confident AI responses may contain errors and fabricated sources
2	Separate claims from evidence	Identify which statements require verification
3	Verify references	Check titles, authors, journals and DOIs in PubMed, Crossref and publisher records
4	Assess relevance	Read the original source and decide whether it supports the claim
5	Correct and document	Revise inaccurate statements, keep verified references and state uncertainty

Step	Name	Student activity
6	Reflect on responsible use	Consider when AI use is appropriate and how it is disclosed under institutional rules

Source: developed by the author.

The effectiveness of the model should be tested in an intervention study, for example a pre-test/post-test design with a control group, using the instrument described above.

Discussion

The reviewed evidence shows that fabricated and inaccurate references are a well-documented and persistent property of generative AI, not an occasional malfunction. Even with newer models, a meaningful share of generated citations remain unreliable [6; 21]. At the same time, the few available studies of human performance suggest that recognising AI errors is difficult even for trainees with professional education [22]. Taken together, this means that responsibility for accuracy cannot be delegated to the tool and must be supported by explicit teaching of verification skills.

For universities in Uzbekistan, the national AI strategy [16] creates both an opportunity and an obligation: the expansion of AI in education should be accompanied by systematic development of AI literacy. Practical recommendations include integrating source verification into academic skills courses; teaching students to use PubMed, Crossref and publisher records; adopting clear institutional rules on acceptable AI use and disclosure; using assessment tasks that require students to justify their reasoning; and training teaching staff to recognise common forms of AI-generated inaccuracy.

Limitations. This is a narrative review rather than a systematic review, and the selection of sources may not be exhaustive. The empirical studies discussed involve different models, tasks and populations, and their results cannot be generalised directly to students in Uzbekistan. The proposed instrument has not yet been validated, and no original data have been collected. Because LLMs are updated frequently, error rates reported for specific model versions may change over time.

Conclusion

1. Generative AI can produce factual errors, unsupported claims and fabricated or inaccurate scientific references in a fluent and convincing form. Empirical studies report that a substantial share of AI-generated references are fabricated or contain errors.
2. Detecting such errors requires two distinct skills — evaluating the accuracy of claims and verifying the existence and relevance of sources — grounded in information literacy, critical appraisal and research methodology.
3. Available evidence suggests that recognising AI hallucinations is difficult even for medical trainees, which underlines the need for targeted instruction.

4. A multicentre cross-sectional study using a validated 24-item instrument is proposed to measure these competencies among university students in Uzbekistan, followed by evaluation of the proposed six-step educational model.

5. The results of such research could inform institutional AI policies and curricula in line with the national Strategy for the Development of Artificial Intelligence Technologies until 2030.

References / Adabiyotlar

1. Association of College and Research Libraries. (2016). *Framework for Information Literacy for Higher Education*. American Library Association. https://www.ala.org/acrl/standards/ilframework

2. 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

3. Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (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

4. Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*, 15(5), e39238. https://doi.org/10.7759/cureus.39238

5. Booth, W. C., Colomb, G. G., Williams, J. M., Bizup, J., & FitzGerald, W. T. (2024). *The Craft of Research* (5th ed.). University of Chicago Press.

6. 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. *Journal of Medical Internet Research*, 26, e53164. https://doi.org/10.2196/53164

7. Creswell, J. W., & Creswell, J. D. (2022). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (6th ed.). SAGE Publications.

8. Gravel, J., D'Amours-Gravel, M., & Osmanlliu, E. (2023). Learning to fake it: Limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3), 226–234. https://doi.org/10.1016/j.mcpdig.2023.05.004

9. Greenhalgh, T., & Dijkstra, P. (2024). *How to Read a Paper: The Basics of Evidence-Based Healthcare* (7th ed.). Wiley. https://doi.org/10.1002/9781394206933

10. 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 Computing Surveys*, 55(12), Article 248, 1–38. https://doi.org/10.1145/3571730

11. Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

12. Kasneci, E., Sessler, K., Küchemann, S., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

13. Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

14. Kung, T. H., Cheatham, M., Medenilla, A., et al. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

15. Miao, F., & Holmes, W. (2023). *Guidance for Generative AI in Education and Research*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

16. O'zbekiston Respublikasi Prezidentining 2024-yil 14-oktabrdagi PQ-358-son "2030-yilgacha sun'iy intellekt texnologiyalarini rivojlantirish strategiyasini tasdiqlash to'g'risida"gi qarori. https://lex.uz/docs/7159258

17. Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

18. Pratama, H. (2025). Training students to identify and correct fabricated references in ChatGPT-generated literature reviews. *Conference on English Language Teaching (CELTI) Proceedings*, 25–36. https://doi.org/10.24090/celti.2025.1335

19. Russell, S. J., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

20. Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11(6), 887. https://doi.org/10.3390/healthcare11060887

21. Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, 14045. https://doi.org/10.1038/s41598-023-41032-5

22. Zhou, J., Zhang, J., Wan, R., Cui, X., Liu, Q., Guo, H., et al. (2025). Integrating AI into clinical education: Evaluating general practice trainees' proficiency in distinguishing AI-generated hallucinations and impacting factors. *BMC Medical Education*, 25, Article 406. https://doi.org/10.1186/s12909-025-06916-2