Al Ain University
College of Business

Scientific Research Methodology [0508203]

Fall 2025- 2026

AI in Academic Research: Assessing the Credibility and Performance of General-Purpose and Research-Focused Systems

By:

Student Name	ID #
Mahmoud Husam Abdeen	202211625

Submitted to: Dr. Maha Rahrouh

Abstract

This paper presents a comparison of the reliability and academic usefulness of general-purpose AI tools (including ChatGPT, Gemini and Copilot) as against research-oriented AI systems (including Consensus and perplexity) in creating correct and verifiable academic sources. The study is interested in analyzing what kind of AI is more accurate in citing, finding universal patterns of error, and setting principles of responsible use of AI in research in higher education. The experimental design will be comparative in collecting and analyzing 20 AI-generated responses in five tools (4 prompt to each ai) Hallucination rates and time of verification will be measured using quantitative data and feedback obtained regarding the ease of verification, conducted using qualitative data. The references will be verified using (Scopus, DOAJ, PubMed, or Google Scholar). Many consider that AI tools centered on research are superior, but all AI tools must be revised to confirm academic integrity. This study can contribute to the creation of ethical and practical principles of responsible use of AI in academics. The research will help to create some ethical and practical principles of the responsible use of AI in academic studies. It is expected that Research-Focus AI tools are likely to produce less hallucinated or wrong references than general-purpose AI systems. Also, the general-purpose AI models will show increased productivity and reduced verifiability.

Contents

Title: AI in Academic Research: Assessing the Credibility and Performance of General-Purpose and Research-Focused Systems

1-Introduction

1.1Background:

The use of artificial intelligence (AI) has become an inseparable part of contemporary academic study. It helps with writing, summarizing, and analyzing data, which saves much time for the students and researchers. General-purpose AI such as ChatGPT, Gemini/Bard, and Copilot, are used to draft, brainstorm, and make referrals. Regardless, such tools are susceptible to issues of inaccuracy, falsified sources, and untestable quotes that may compromise the academic standards and reliability of research.

On the other hand, AI tools, including Consensus and Perplexity, are research-based, implying that they are linked to peer-reviewed and transparent and verifiable databases. They focus on accuracy and evidence-based products, and state to provide a possibly improved alternative to literature reviews and evidence synthesis.

Considering these differing abilities and constraints, the present proposal aims to investigate the scholarly worth of AI technologies on record through the prism of analyzing the distinctions between general-purpose tools and research-based platforms concerning their scales of accuracy, reliability, and verifiable transparency.

1.2 Motivation:

The increased use of AI tools in the research has brought up ethical and academic issues. Although they are useful, AI hallucinations, in which a system will create false or wrong information, continue to be a significant problem. Several studies have pointed out that the practice of fabricated quotations and false DOIs is common, causing incorrect information and poor research results (Chelli et al., 2024; Hueber and Kleyer, 2023).

Identifying which AI tools are more reliable will help researchers and students make informed decisions about when and how to use them. This paper seeks to address the gap in technological innovation and academic integrity by looking at the effectiveness of AI and establishing best practices that can be used to ensure responsible application in higher education.

1.3 Research Aim(s):

The aim of this study is to compare the general-purpose AI tools (such ChatGPT, Gemini, Copilot) and research-focused AI tools (such Consensus and Perplexity) to determine which type generates more accurate and real academic references, which ones are easier and faster to verify and which AI tool can produce a clearer, more accurate, and academically appropriate research paper. This comparison will help students to understand how to use AI tools safely, responsibly and effectively in university-level research.

1.4 Research Question(s):

- What kind of AI, a general-purpose or research-Focus one, generates more accurate and verifiable references?
- What are the most frequent mistakes (i.e. hallucinations, wrong metadata) of AI-generated references?
- Which strategies are the best practices to minimize citation errors in the use of AI for academic purposes?

1.5 Research Objectives

- Compare the accuracy of academic referencing tools created by general-purpose AI tools vs. research-focused AI tools.
- To detect the common citations errors committed by each type of AI, like fake references, wrong authors, wrong years, and wrong DOIs or links.
- To test how well each AI tool can write a complete research paper, in regard to correctness of words, structure, and references being used correctly.
- To create easy to understand and implement guidelines for students on ways to use AI tools safely and responsibly in academic research.

"In other to achieve the above declined objectives, we will critically review the literature in section 2, followed by a presentation of our adopted methodology in section 3. Ethical considerations and procedures will be discussed in section 4, the time scale for our work will be provided in section 5, while section 6 will describe the needed resources, 7 Appendices, finally will list the references used in the project."

2-Literature-Review:

2.1 Scope and approach

This is a literature review on the efficacy of AI tools in scholarly research that focuses on the accuracy of citation, hallucination, and verifiability. Our general-purpose chatbots are broken down into smaller (and more specific) in relation to the underlying research problem: the general-purpose chatbots (e.g., ChatGPT, Gemini/Bard and Copilot) versus the research-focused or domain-tethered tools (e.g., Elicit, Consensus and perplexity) We generalize what is familiar, find gaps and unanswered issues, record areas of disagreement, and describe theoretical perspectives and research approaches that have been implemented so far.

2.2 What is already known?

AI increases productivity in research activities (e.g., drafting, scanning literature) and results are not as productive or reliable depending on tasks and tools. Empirical analyses demonstrate that ChatGPT can effectively organize academic texts with the help of the drafts, but the literature-review and methodology sections may need a significant amount of revision and verification, which makes the question of authorship and integrity highly controversial. (Khlaif et al., 2023)

The relevance of the general-purpose chatbots is often that they tend to hallucinate references (what they create or cannot verify) and to mis-tag bibliographic information (authors, titles, year, volume/issue, DOI) core, well-documented finding. Benchmarking systematic-review prompts revealed high rates of hallucinations of ChatGPT and Bard/Gemini with weak recall and GPT-4 dominated over GPT-3.5 but

was still far below human curation levels. (Chelli et al., 2024)

Large cross-topic tests also give fabricated results with high error rates: in one Scientific Reports study, most of the non-fabricated citations still had bibliographic errors. These findings are universal, and they apply in other areas, such as in clinical contexts where false references have been reported in actual practice. (Walters & Wilder, 2023)

Numerous studies indicate that generative AI is an important boost in productivity in academic and professional tasks. AI tools help people to accomplish writing, summarizing and data processing tasks more efficiently thereby decreasing workload and time consumption of complex tasks. Findings suggest that general-purpose AI can be used for speeding up early-stage research tasks like drafting and brainstorming that can be especially appealing for students and researchers looking for efficiency. However, these improvements in speed are frequently paid off with respect to reliability and verifiability. (Al Naqbi et al., 2024)

2.3 General-purpose vs. research-focused tools

Law evidence demonstrates that databased, research-oriented tools decrease hallucinations compared to general-purpose LLMs, but do not eliminate them: an empirical test of Lexis+ AI, Westlaw AI-Assisted Research, and Ask Practical Law AI preregistered found hallucinations with between 17 and 33% accuracy, although vendors claim their systems have no hallucinations at all. This proves the hypothesis that grounding leads to better reliability as well as proving the necessity to verify even specialized systems. (Magesh et al., 2025).

The tool choice is associated with mixed effects on factual accuracy and readability, with respect to medical comparison of ChatGPT-4o, Copilot and Perplexity (e.g., Perplexity higher factual

accuracy; Copilot more readable) suggesting that the effectiveness is multi-dimensional (accuracy, clarity, verifiability). (Zorlu & Günal et al., 2025)

2.4 Unanswered questions and gaps

In spite of the rapid progress there are important gaps:

- There is a relative dearth of direct and cross-task comparisons of general purpose and research-oriented tools when not within particular fields (e.g., law, medicine). In most studies, it is one task or domain that is tested (not both academic tasks: search, summary-with-citations, drafting with in-text references). (Chelli et al., 2024)
- We do not have fine-grained condition-specific error typologies (e.g. prompt complexity, task type, domain) - although emerging work has suggested structured scores of reference authenticity between chatbots and exhibits complexity effects. (Aljamaan et al., 2024)
- Minimum quantification of the time/effort to check AI-generated references (a key aspect of performance required by students and researchers) is done. Recent studies focus on the prevalence of errors rather than on verification costs. (Walters & Wilder, 2023)
- While the issue of hallucination in AI output is well known, there are large gaps in understanding its underlying causes. Most existing research contains examples of the content of hallucinations but fails to fully explain why these errors occur more often in some tasks than others. Jamaluddin says that the field lacks a coherent conceptual framework to explain the mechanisms for fabricated outputs that inhibits efforts to predict or avoid such errors. As a result, the study of hallucinations remains in the descriptive rather than diagnostic stage and major theoretical gaps can persist. (Jamaluddin, 2023)

2.5 Areas of professional conflict

There are two tensions which run through literature. To start, productivity vs. reliability: general purpose tools will faster write results than research focus tools (can accomplish verifiability); research focus tools (usually offer better checks) are able to accomplish. Second, measurement and definitions of hallucination: although research operationalizes the term hallucination (fabrication vs. serious metadata errors) there are cross-study comparisons, so it may be time to standardize and transparent evaluation frameworks. (Bang et al., 2023)

A large professional conflict ensues because of the disconnect between what AI developers say and how systems actually perform. While the research-Focus AI technology tools of a few vendors claim to be "hallucination-free", empirical evaluations refute this. Comparative audits of legal research AI systems revealed that there is between 17% and 33% of hallucination rate in spite of being promoted as fully reliable, meaning that even domain-specific AI continues to be plagued by accuracy problems. This discrepancy is part of continuous debates on the trustworthiness of AI and importance of the ethical aspects of climbing models that tend to overstate their capabilities. (Magesh et al., 2025)

2.6 Theories and conceptual perspectives

Credibility and efficiency depend on openness and foundation. NLG literature puts hallucination in the frame of a systemic property of generative models and is a list of mitigation measures (e.g., retrieval-augmented generation, verification). This theoretical framework contributes to our theoretical model: Tool Type (Grounding/Transparency) Reference Quality with moderator variables being Task Type and Prompt Complexity. (Bang et al., 2023)

2.7 Suggestions for further research

The standards suggested by the field include:

- uniform standards of reference precision/ recall inter-tool.
- cost-of-verification.
- pedagogical directions of responsible usage. There are a number of such studies which start to formalize the scoring (e.g.: Reference Hallucination Score), but still comprehensive, cross-disciplinary standards of adoption and reporting are required. (Aljamaan et al., 2024)

- Recent evaluations on AI-generated responses in specialized fields highlight the need for better frameworks of accurate assessments. Research indicates that chatbot responses differ greatly in quality and frequently contain misleading information or partially correct information. Shiferaw and colleagues point out the need for creating standardized benchmarks by which to assess AI output reliability, accuracy and consistency in various domains. Their work has implications for future research that should focus on developing rigorous methods for testing in order to mitigate the risks of misinformation and enhance the accountability of the models. (Shiferaw et al., 2024)

2.8 Research strategies used in prior work

Controlled experiment (systematic prompts, blinded rating), domain specific audit (medicine, law) and large-scale citation audit (thousands of references) are all present in the corpus. Some of these methods are both manual verification of publisher records/DOI registries and inter-rater reliability

as well as statistical comparisons across models and tasks which we follow and expand our methodology. (Chelli et al., 2024)

3-Methodology

3.1 Research Design

An experimental design of the research is comparative based on the combination of quantitative and qualitative methods.

Experimental comparative designs are widely adopted for testing the differences in accuracy, performance and error rates of AI systems because these designs enable researchers to measure outcomes under controlled and repeated conditions. (Ferri et al., 2009)

- **Quantitative**: Determine the quality of citations, frequency of hallucinated references or fabricated references and potentially the percentage accuracy of verified sources on general-purpose and researcher-oriented AI tools.
- **Qualitative**: Evaluate the quality of wording, the structure, and the general accessibility of the AI-generated academic products, the nature and frequency of errors in citation.

A mixed-methods approach is recommended if we want to study the behavior of AI because the former provides quantitative measures of performance results while the latter offers qualitative feedback on user experience, interpretation and difficulty of verification.

With the help of this design, it is possible to compare the two types of AI systems under the same conditions in a systematic and fair way. It assists in determining quantifiable shifts in performance, finding common patterns of citation acts, and determining the dependability and efficacy of every tool in generating scholarly valid research materials.

3.2 Data collection procedure:

The data collection process was meant to evaluate the research performance, accuracy, and reliability of various AI systems. All the AIs were to complete four structured tasks, which are

1. Find 50 peer-reviewed academic references with link concerning [Title].
2. Write the references in APA 7th edition and list them alphabetically.
3. Write full research about [Title] with reference.
4. Summarize this research [non-existent research].

The four tasks that were assigned to AI systems were done in the same conditions. The results of all outputs were recorded and assessed in table form with the help of a standardized rubric. The rubric evaluated the reference validity, source checking, accuracy of citation, APA style, comprehensibility of the text, and research completeness. The performance of each AI was also summed based on the mean score of the four tasks.

In addition to the main tasks, the AI tools were asked to summarize a non-existent study attributed to a fictitious author.

The references produced by the AIs were verified in terms of authenticity and peer-reviewed status by using the trusted academic indexing databases (Scopus, DOAJ, PubMed, and Google Scholar) and, in some cases, other verification tools. Such systematic and comparative methodology allowed providing fair evidence-based consideration that could show which AI system has the best overall research potential.

3.3 The Theoretical / Conceptual Framework:

The conceptual model was informed by the reviewed literature: Tool Type (general-purpose vs. research-focused) and mediated by Grounding Transparency, and moderated by Task Type, affect Citation Accuracy.

- Independent variable: Type of AI Tool
- Intervening variables: Verifiable Academic Database Integration.
- Dependent Variables: Citation Accuracy, Hallucination rate, verifying Time.
- Extraneous: Task Type

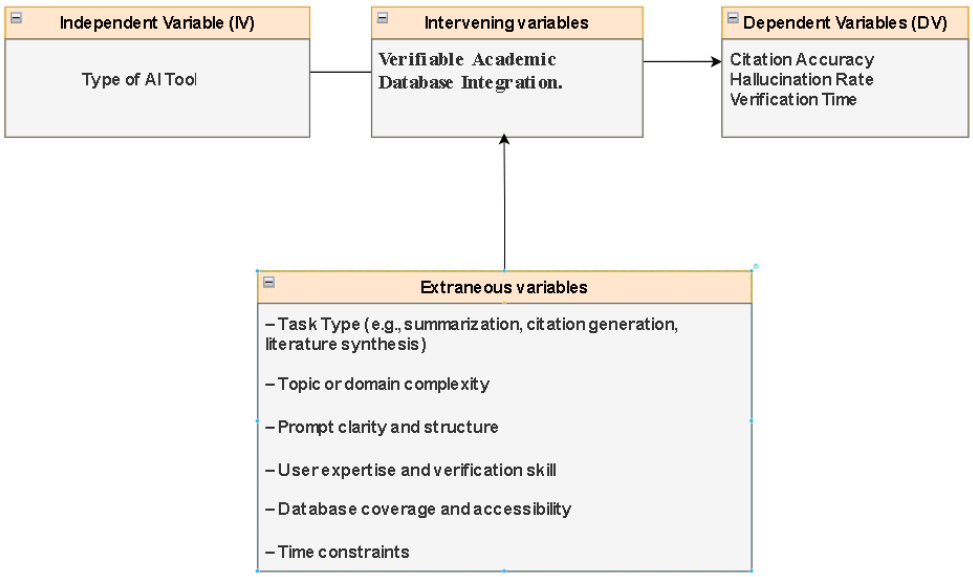

Figure 1. Conceptual Framework of the Study (Created in draw.io)

The theoretical framework as depicted in the figure 1 shows the association between the main variables that are studied in the current research. The type of AI tool is the independent variable that affects the outcomes by means of one of the intervening variables the level of the integration with verifiable academic databases. The intervening variable should have a mediating effect on the dependent variables, namely citation accuracy, and hallucination rate and, finally, verification time. Also, the framework takes into consideration various extraneous factors that can have an impact, such as type of task, complexity of topic, clarity of prompt, user expertise, coverage of the databases, and time limitation. All of these elements prove that the study is a systematic and contextually relevant performance evaluation of AI.

3.4 Data Analysis:

The standardized evaluation rubric that will be used to analyze all outputs created by the five AI tools systematically is the one that is presented in Appendix B. There are various criteria in this rubric, including checking references, compliance with APA 7th formatting, metadata compliance, resolving DOIs, the percentage of plagiarism, the correctness of facts, and writing clarity, as well as completeness of the research overall. The AI responses will be rated based on these and each of the elements will be judged 100 percent.

After rubric-based assessment, all the sources mentioned by the AI systems will be checked separately with the help of CrossRef, Google Scholar, Scopus, and other well-known indexing databases. The number of genuine references and the amount of fake, unprovable or misleading references will be counted on each tool. These numbers will be then used to obtain the percentage of correct references and the hallucination rate of each AI system. Besides, the duration of checking each reference will be timed to determine efficiency.

All findings will be tabulated in relative form such as rubric scores, accuracy of verification, the rates of hallucinations, and time figures. These tables will allow having a clear performance comparison between general AI tools and research-oriented AI systems. Lastly, conclusions will be drawn on the aggregated findings on which type of AI possesses superior credibility, accuracy and verifiability in the creation of academic content.

4-Ethical considerations and procedures:

As a way of being just and avoiding bias, I will provide the same prompts to all AI tools in the same circumstances. Each tool will have a single attempt and no second or edited many attempts because it is necessary to maintain consistency in results. No tool will be preferred or altered in its output, and no data will be rated by some different verification checklist and criteria. This analysis will not refer to the personal opinions and preferences but only to the accuracy and verifiability of the references. The AI tools will be handled in the same way to make sure that the findings are objective, unbiased, and reliable.

5-Timescale:

The Gantt chart depicts the key activities of the research between November 2025 and March 2026, with literature review and final submission being the first and the last activity respectively. The tasks will be planned week by week to maintain track of the project. The chart is prepared for Microsoft excel.

Month		Nov 2025				Dec 2025				Jan 2026				Feb 2026				Mar 2026			
Activity	Week number	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18	W19	W20
Read literature	1																				
Finalize objectives	2																				
Draft literature review	3																				
Read methodology literature	4																				
Devise research approach	5																				
Draft research proposal	6																				
Develop questionnaire	7																				
Pilot test & revise questionnaire	8																				
Administer questionnaire	9																				
Enter data into computer	10																				
Analyze data	11																				
Draft findings chapter	12																				
Update literature	13																				
Complete remaining chapters	14																				
Submit to tutor	15																				
Review draft & format	16																				
Print & bind	17																				
Submit	18																				
Reflective diary	19																				

Figure 2: Project Gantt Chart Showing Research Activities from November 2025 to March 2026 (Created in Microsoft Excel)

6-Resources

- Access to AI (ChatGPT and perplexity).
- Internet databases (CrossRef, Google Scholar, Scopus).
- Statistical software (Microsoft Excel).
- Course instructor
- Software for verification (CrossRef, PubMed, Scopus)

7-Appendices:

7.1 Appendix A: Sample AI Task Prompts:

1. Find 50 peer-reviewed academic references with link concerning [Title]
2. Write the references in APA 7th edition and list them alphabetically.
3. Write full research about [Title] with reference
4. Summarize this research [non-existent research].

The AI tools are to be tested with the help of four tasks. To begin with, every AI will be requested to identify 50 peer-reviewed academic references associated with a particular topic, and in this way, we will be able to evaluate accuracy and authenticity of the sources provided by the AI. Second, the AI will cite these sources using the correct APA 7th edition style and organize them in alphabetical order to determine whether the AI can adhere to the rules of academic citation. Third, both AI will produce complete research paper on the chosen topic, which will allow us to assess the clarity and structure of academic writing, and the quality of its referenced sources. Lastly, the AI will have to sum up a fictional research study by a fictitious author, which will be meant to evaluate whether the system can identify false information or the system is generating a hallucination. All the output would be saved.

7.2 Appendix B: Rubric table to evaluate each response:

Criterion	Evaluation steps	Score	Weight (%)
Checking references in accepted academic indexing databases.	Whether every source is available at Scopus, DOAJ, PubMed, or Google Scholar.	Number of available references: ____ / 50	15%
APA 7th (alphabetical)	Ensure that all references are following APA 7th regulations and are arranged in alphabetical order.	- Is it in APA 7th Style? - its alphabetical order?	15%
The author information, year, and journal information are the same	Make sure that the name of the author, year and the details of the journal used are the same as the official record.	Number of references fully correct: ____ / 50	15%
DOI or link resolves correctly	Click on every DOI or link to ensure that it links to the right article.	Number of valid link: ____ / 50	10%
Plagiarism	Finding the percentage of plagiarism with a plagiarism checker. https://www.duplichecker.com/	The percentage of plagiarism (A lower plagiarism percentage is better.)	10%

Writing clarity & structure	Ensure the writing is understandable, well structured, and no errors or fantasy information.	The research depends on clear writing and a well-organized structure.	10%
Quality and completeness of the full research	Make sure that the study consists of all key points and that the references are credible and used correctly.	The research depends on clear key points and accurate references.	25%
Overall	Calculate the Average for all Criteria score: - 90–100% = Excellent - 60–80% = Good - 40–60% = Fair - Below 40% = Poor	-	__%

7.3 Appendix C: Guidelines for Safe and Responsible Use of AI in Academic Research:

Artificial intelligence tools can be useful to students when it comes to writing, research and organizing ideas, but they can also make mistakes, present false information or citations that do not exist. Because of this, it is important that students know how to use AI safely and responsibly in their academic work. The guidelines in Appendix C have been developed by considering the findings of this study to offer clear and practical guidance to help the student use AI appropriately and avoid any issues arising from a lack of accuracy, plagiarism, or academic integrity.

1. **Check all the AI-generated references in the academic databases:**

It is necessary to check all the citations in Scopus, Google Scholar, PubMed, DOAJ or the publisher's website in order to make sure that they actually exist.

2. **The misuse of AI-generated text in academic writing:**

Avoid copying AI-generated text into academic writing. AI output should be rewritten, evaluated or critically adapted so that the originality and avoiding plagiarism are preserved.

3. **As AI-generated content is used, it is important to remember that the content is not final writing but a draft:**

AI should not be used to replace critical thinking and academic argumentation, but to support the generation or structuring of ideas.

4. **Do not trust AI with research methodology, data collection or results:**

Students are required to create authentic approaches, gather authentic data, and independently analyze data.

5. Check all the claims that A.I. generates for hallucinations:

Study that is fabricated, facts that are incorrect or statements that cannot be verified must be ruled out of academic work.

6. Use AI in an ethical manner and do not rely on it to complete the entire assignments:

AI should be a way of enhancing learning, not a replacement for student effort and violating academic regulations.

7. Safeguard personal and academics information upon using AI instruments:

Students should not enter into personal information or sensitive research content into third-party systems.

References

1- **Aljamaan, F., Alqurashi, F., Alalawi, Y., Alhazmi, F., & Aljohani, N. (2024).** Reference hallucination score for medical AI chatbots: Development and usability study. *JMIR Medical Informatics, 12* (1), e25115. https://pmc.ncbi.nlm.nih.gov/articles/PMC11325115/

2- **Al Naqbi, W., Almarzooqi, A., Almarzooqi, S., & Al Shamsi, F. (2024).** Enhancing work productivity through generative artificial intelligence. *Sustainability, 16* (3), 1166. https://doi.org/10.3390/su16031166

3- **Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023).** *A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity* **(arXiv Preprint No. 2302.04023). arXiv.** https://arxiv.org/abs/2302.04023

4- **Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., & Ruetsch-Chelli, C. (2024).** Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research, 26* (1), e53164. https://doi.org/10.2196/53164

5- **Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009).** An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(27-38). https://doi.org/10.1016/j.patrec.2008.08.010

6- **Hueber, A. J., & Kleyer, A. (2023).** Quality of citation data using the natural language processing tool ChatGPT in rheumatology: Creation of false references. *RMD Open, 9* (2), e003248. https://doi.org/10.1136/rmdopen-2023-003248

7- **Jamaluddin, N. (2023).** Hallucination: A key challenge to artificial intelligence-based systems. *Cureus, 15* (7), e42170. https://pmc.ncbi.nlm.nih.gov/articles/PMC10726751/

8- **Khlaif, Z. N., Mousa, A., Hattab, M. K., Itmazi, J., Hassan, A. A., Sanmugam, M., & Ayyoub, A. (2023).** The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. *JMIR Medical Education, 9* (1), e47049. https://doi.org/10.2196/47049

9- **Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025).** Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies, 22* (2), 216–242. https://doi.org/10.1111/jels.12413

10- **Shiferaw, M. W., Zheng, T., Zhang, X., & Zhou, L. (2024).** Assessing the accuracy and quality of AI chatbot-generated responses to healthcare inquiries. *BMC Medical Informatics and Decision Making, 24* (1), 182. https://doi.org/10.1186/s12911-024-02824-5

11- **Walters, W. H., & Wilder, E. I. (2023).** Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13* (1), 16118. https://doi.org/10.1038/s41598-023-41032-5

12- **Zorlu, Ö., Günal, U., & Yazici, S. (2025).** The quality, accuracy, and readability of information about hidradenitis suppurativa provided by artificial intelligence: Comparative analysis of ChatGPT, Copilot, and Perplexity. *Medicine, 104* (39), e44728. https://doi.org/10.1097/MD.0000000000044728