ORIGINAL PAPER

Evaluation of AI Citation Accuracy in Anterior Segment Research

Civelekler Mustafa, Çıtırık Mehmet

University of Health Sciences, Ankara Etlik City Hospital, Department of Ophthalmology, Ankara, Türkiye

Mustafa Civelekler, MD

Correspondence address:
University of Health Sciences, Ankara Etlik City Hospital, Department of Ophthalmology
Ankara
Türkiye
E-mail: mcivelekler@yahoo.com

Submitted to the editorial board: December 23, 2025
Accepted for publication: March 13, 2026
Available on-line: May 7, 2026

SUMMARY

Aims: To conduct a pilot evaluation of the citation accuracy of four contemporary artificial intelligence (AI) models – ChatGPT (OpenAI GPT-5.1), Copilot (Microsoft Copilot 4.2), DeepSeek (DeepSeek-R1), and Gemini (Google Gemini Ultra 2.5) – in generating PubMed-style references for corneal, conjunctival, and eyelid disease research, and to identify common error patterns.

Material and Methods: Thirty-five standardized clinical paragraphs were selected from *The Review of Ophthalmology* (4th edition). Each AI model was prompted to generate AMA 11-style references relevant to the provided text, simulating a literature retrieval task. Generated citations were assessed for accuracy, DOI matching, and clinical relevance. In a second validation phase, citations were independently reviewed by two ophthalmology experts and classified as fully cited, partially cited, or not cited. Statistical comparisons of accuracy proportions among models were performed using chi-squared tests.

Results: DeepSeek demonstrated the highest citation accuracy (78.6%, 22/35), followed by ChatGPT (51.4%, 18/35), and Copilot (51.4%, 18/35). Gemini showed the lowest accuracy (12.9%, 5/35). Differences in accuracy rates across models were statistically significant ($\chi^2 = 19.0$, df = 3, $p < 0.001$). Expert validation confirmed DeepSeek's relative advantage, with 42.9% (15/35) of its references classified as fully cited, compared with Copilot (20.0%, 7/35), ChatGPT (11.4%, 4/35), and Gemini (11.4%, 4/35). The most frequent error types were DOI mismatches and the generation of irrelevant or unverifiable references.

Conclusion: This pilot study indicates that contemporary AI models, particularly those like DeepSeek, show potential in assisting with citation generation. However, the observed error rates, including instances of hallucination, remain substantial. These findings underscore that rigorous human verification is indispensable when using AI for academic referencing in specialized medical fields, and highlight the need for continuous, version-specific benchmarking as these tools evolve.

Key words: artificial intelligence; citation accuracy; corneal disease; conjunctival disorders; eyelid diseases; large language models

Čes. a slov. Oftal., 82, 2026, No. x, p.

INTRODUCTION

The integration of artificial intelligence (AI) into healthcare continues to expand, with demonstrated potential in areas ranging from diagnosing diabetic retinopathy to managing glaucoma [1,2]. Concurrently, AI's role in augmenting academic workflows, including the management of scholarly references, is gaining attention [3]. AI-powered tools could potentially streamline literature reviews and ensure citation consistency. However, as these technologies are rapidly adopted, fundamental questions regarding their reliability in specialized academic contexts require careful, empirical investigation.

Accurate citation is a cornerstone of scientific integrity, ensuring the traceability and credibility of research. While established reference managers, such as EndNote and Zotero, provide systematic control, the allure of AI lies in its potential to reduce manual effort. A significant barrier is the phenomenon of "AI hallucinations", where models generate convincing, but entirely fabricated references [4,5]. Such errors, stemming from limitations in training data, pose a particular risk in specialized fields such as ophthalmology, where incorrect citations could misdirect research and clinical understanding.

Several publicly accessible AI models are increasingly used for academic tasks. This pilot study selected four contemporary and widely discussed models – ChatGPT (OpenAI GPT-5.1), Copilot (Microsoft Copilot 4.2), DeepSeek (DeepSeek-R1), and Gemini (Google Gemini Ultra 2.5) – to conduct an initial, comparative exploration in late 2025. While ChatGPT and Copilot are versatile general-purpose models, DeepSeek has been noted in prior

reports for its relatively stronger performance on biomedical tasks [6]. Gemini represents a newer, advanced general-purpose architecture. Despite their growing use, a systematic, head-to-head assessment of their ability to generate accurate citations within a focused medical subfield was lacking.

This exploratory study aimed to evaluate and compare the effectiveness of these four AI models in generating PubMed-style citations, specifically for anterior segment disorders (corneal, conjunctival, and eyelid diseases). This focused scope was chosen to allow for a controlled, in-depth analysis of error patterns within a coherent domain. The primary objective was to assess the baseline citation accuracy of each model when provided with a standardized clinical text. A secondary goal was to categorize prevalent error types. By detailing the strengths and limitations observed in this pilot evaluation, we aim to inform researchers about the current capabilities of these tools and to underscore the critical importance of human oversight. The findings also provide a foundation for future, more comprehensive studies across other ophthalmic subspecialties.

MATERIAL AND METHODS

This pilot study aimed to conduct an initial assessment of the citation generation performance of four contemporary artificial intelligence (AI) models within the domain of anterior segment ophthalmology. All models were accessed and evaluated over a defined period between November 15 and December 20, 2025. The specific model versions tested were: ChatGPT (OpenAI; GPT-5.1 architecture, publicly accessible version at the time of access), Copilot (Microsoft Copilot 4.2), DeepSeek (DeepSeek-R1), and Gemini (Google Gemini Ultra 2.5). Their performance reflects the capabilities of these specific iterations at that point in time.

Models were accessed via their publicly available web interfaces, using default inference settings; no temperature or system-level parameters were modified. Publicly disclosed training cut-offs, where available in vendor documentation, were recorded as contextual information.

The models were selected to represent a mix of general-purpose and medically-oriented tools available in late 2025. ChatGPT (GPT-5.1) and Copilot (4.2) are broad-based, multi-modal models not specifically trained for bibliographic tasks. DeepSeek-R1 was included due to emerging reports suggesting its proficiency with medical literature [6]. Gemini Ultra 2.5 was included as a representative of newer, advanced general-purpose models optimized for complex reasoning.

To create a standardized testbed, 35 clinical paragraphs covering distinct corneal, conjunctival, and eyelid pathologies were sourced from *The Review of Ophthalmology* (4th edition). This source provided consistent, textbook-style summaries. Paragraphs were selected to be self-contained and representative of typical clinical descriptions. While this sample size permits an initial controlled comparison, it is acknowledged as a limited set suitable for a pilot investigation.

Paragraphs were selected by consensus of the two authors to represent common, well-described anterior segment conditions and to ensure balanced coverage across corneal, conjunctival, and eyelid disorders.

A fixed, single-round prompt was used for consistency: "Please provide an AMA 11-style reference for sources relevant to the following clinical paragraph." It is crucial to clarify that this instruction was designed to simulate a common user request for citation assistance and to probe the models' internal knowledge bases. No model had real-time access to PubMed or other databases during this evaluation; all outputs were generated from their pre-trained knowledge. This design choice was intentional to evaluate baseline performance without the confounding variable of live data retrieval.

Retrieval-augmented generation (RAG) and any real-time database access were intentionally excluded to isolate baseline citation behavior from each model's pre-trained knowledge.

Generated citations were evaluated against pre-defined criteria:

1) Correct Publications (verifiable and relevant),
2) DOI Mismatches (correct journal, wrong DOI),
3) Incorrect Author/Journal Name, and
4) Non-Correct Publications (unrelated or unfindable).

The primary outcome was the proportion of correctly generated citations per model.

Subsequently, all citations initially flagged as "correct" underwent expert validation. Two ophthalmology experts independently classified each citation as: "Fully Cited" (exact match), "Partially Cited" (relevant but not the best or exact match), or "Not Cited" (unrelated or fabricated). Discrepancies were resolved by discussion. Inter-rater reliability for this human evaluation was substantial (Cohen's kappa = 0.74). During manuscript preparation, ChatGPT Plus was used only as an auxiliary tool to verify DOI formats, not for primary analysis.

Statistical Analysis

Descriptive statistics summarized the performance of each model. Given that the primary outcome was a proportion (citation accuracy rate), comparisons across the four models were performed using chi-squared (χ^2) tests for independence, to determine if the distribution of correct vs. incorrect citations differed significantly. This approach is appropriate for comparing categorical outcomes across multiple groups. A secondary chi-squared test was used to analyze the distribution of error types (DOI Mismatches, Journal/Author Mistakes, Non-Correct Citations) across models. All analyses were performed using SPSS version 23 (SPSS Inc., Chicago, IL, USA).

This study utilized publicly available AI tools and the PubMed database. It did not involve human subjects, personal data, or the reproduction of protected content, thus ethical approval was not required.

RESULTS

The performance of the four AI models in generating citations for anterior segment disorders is summarized in Table 1. Citation accuracy varied substantially across the models. DeepSeek achieved the highest proportion of correct citations at 78.6% (22/35). ChatGPT and Copilot demonstrated moderate and similar performance, with accuracy rates of 51.4% (18/35) and 51.4% (18/35), respectively. Gemini exhibited the lowest accuracy, at 12.9% (5/35). A chi-squared test confirmed that these differences in accuracy distribution were statistically significant ($\chi^2 = 19.0$, df = 3, $p < 0.001$).

Graph 1 provides a visual comparison of the error profiles. A detailed breakdown reveals distinct patterns: DeepSeek had the highest number of DOI mismatches (8 instances), while producing the fewest non-correct citations (13 instances). In contrast, Gemini generated the highest volume of non-correct citations (31 instances), and no journal/author name mistakes. Copilot was unique in having zero DOI mismatches, but produced 17 non-correct citations. The distribution of these three error categories (DOI Mismatches, Journal/Author Mistakes, Non-Correct Citations) across the four models was statistically significant ($\chi^2 = 17.8$, df = 6, $p = 0.007$).

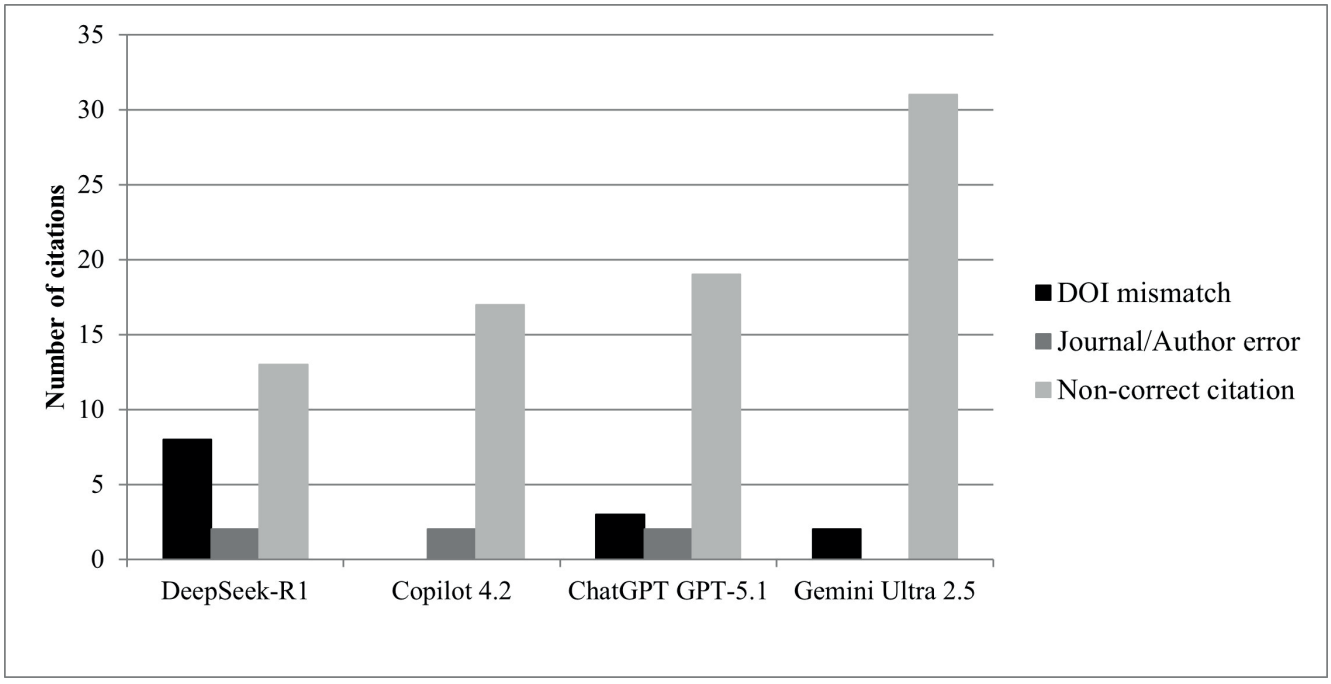
The expert validation phase provided a stricter assessment of citation quality. DeepSeek yielded the highest number of references classified as "fully cited" by experts (15 out of 35, 42.9%). However, a notable portion of its output was still deemed only "partially cited" (12 instances) or "not cited" (5 instances). The other models had lower expert-validated fully-cited rates: Copilot (20.0%, 7/35), ChatGPT (11.4%, 4/35), and Gemini (11.4%, 4/35). This phase highlights that a significant proportion of AI-generated citations require substantial correction or replacement, even from the best-performing model in this evaluation.

In summary, within the constraints of this pilot study, DeepSeek demonstrated a relative advantage in citation accuracy for anterior segment content. All models, however, were prone to significant errors, with DeepSeek showing a propensity for DOI inaccuracies and the other models, especially Gemini, frequently generating irrelevant or unverifiable citations (Graph 1). The expert review crucially confirmed that human verification remains essential to identify and correct these deficiencies.

Table 1. Performance metrics of artificial intelligence models in citation generation

Model	Correct Citations (%)	DOI Mismatches	Journal/Author Name Mistakes	Non-Correct Citations
DeepSeek	78.6	8	2	13
Copilot	51.4	0	2	17
ChatGPT	51.4	3	2	19
Gemini	12.9	2	0	31

AI – artificial intelligence, DOI – digital object identifier



Graph 1. Comparative distribution of citation error types across four artificial intelligence models
AI – artificial intelligence, DOI – digital object identifier

DISCUSSION

This pilot study conducted an exploratory comparison of four contemporary AI models (late-2025 versions), tasked with generating academic citations for anterior segment ophthalmology literature. The results reveal pronounced variability in performance. DeepSeek-R1 achieved the highest initial accuracy rate (78.6%), while the general-purpose models, ChatGPT (GPT-5.1) and Copilot (4.2), showed moderate performance (~51%), and Gemini Ultra 2.5 performed poorly (12.9%). It is important to interpret the 78.6% accuracy not as a marker of reliability, but as a comparative baseline indicating potential; an error rate exceeding 20% remains clinically and academically significant. The superior performance of DeepSeek is consistent with suggestions that its training may include a stronger representation of biomedical text, though it is not a formally domain-specific tool [6,7].

The analysis of error patterns, visualized in Graph 1, is particularly instructive. The high frequency of "Non-Correct Citations" – essentially hallucinations where the model invents or misattributes references – represents the most critical failure mode, especially for Gemini. This risk is acute in medical research, where citation integrity is paramount [5,8]. DOI mismatches, most common with DeepSeek, while perhaps less severe, still impede efficient access to source material. These errors probably stem from the models' static training data, which may lack paywalled content or current metadata, leading to plausible approximations [9,10]. The lack of journal/author name errors across models is a positive finding, suggesting core bibliographic elements are often preserved, even when the citation itself is flawed.

A key insight from the expert validation is the distinction between automated accuracy and expert-validated utility. Many citations initially deemed "correct" were later judged by specialists to be only partially relevant or entirely wrong. This underscores that the current operational "reliability" of these tools is low and that human expertise is non-negotiable for quality control. The concept of reliability in this context is better understood as the consistency of human expert judgment (which was substantial, $\kappa = 0.74$), rather than an intrinsic, repeatable property of the AI output itself. Furthermore, an analysis of agreement between models on specific citations was not performed, but represents an important area for future research, to differentiate whether errors are systematic or random.

The findings suggest that the most prudent application of current AI citation tools is within a human-in-the-loop framework [11,12]. AI can serve as a rapid first-pass tool to propose potential references, but its output must be rigorously vetted by the researcher, using trusted databases such as PubMed or Scopus. Integration with traditional reference managers, which provide direct database links and structured fields, could help to mitigate some formatting errors, but does not address the core issue of hallucinated content [13].

In practice, AI-generated citations can serve as a rapid initial shortlist, but each reference should be verified (title, authors, journal, year, and DOI) against authoritative sources (e.g., PubMed/CrossRef) before inclusion in a manuscript.

Limitations and Future Directions

Each model was queried once per paragraph. Because large language models employ stochastic generation, outputs may vary across repeated identical prompts. Our results therefore represent a snapshot of model behavior and do not assess test-retest consistency. Future studies using repeated queries could further characterize intra-model variability.

This work has several important limitations that frame it as a pilot investigation. Firstly, the scope was intentionally limited to 35 paragraphs on anterior segment diseases, to enable a detailed analysis. This necessarily excludes other major ophthalmic subspecialties (e.g., retina, glaucoma, pediatrics). Future studies should expand to these areas, in order to test the generalizability of these findings and identify subspecialty-specific performance patterns.

Secondly, the methodological design tested a single, simple prompt without real-time database access. This was a deliberate choice to assess baseline knowledge, but does not reflect more advanced use cases involving retrieval-augmented generation (RAG) [14]. Subsequent research should evaluate how performance changes with optimized prompts, role-playing instructions, or live PubMed integration. The prompt's instruction was recognized as a simulation of a user query, not a technical capability of the models.

Importantly, RAG-enabled systems are increasingly used in practice and may yield different citation accuracy profiles. For this reason, these findings should not be extrapolated to workflows that include live retrieval.

Thirdly, the rapid evolution of AI technology itself is a fundamental constraint. The models evaluated (ChatGPT GPT-5.1, Copilot 4.2, DeepSeek-R1, Gemini Ultra 2.5) are subject to continuous updates by their developers. Our findings are strictly tied to the model versions, training data cut-offs, and capabilities as of our access period (November-December 2025). Performance metrics, error rates, and even relative rankings could shift with subsequent releases. This underscores the exploratory nature of our study and highlights the critical need for continuous, longitudinal, and version-specific benchmarking of AI tools in academic settings.

Finally, the statistical comparison focused on accuracy proportions and error category distributions. More granular analyses, such as examining agreement between models on specific citations or correlating error rates with paragraph complexity or publication date, were beyond the scope of this initial evaluation, but represent valuable avenues for future research.

CONCLUSION

This exploratory pilot study provides a snapshot evaluation of four contemporary AI models (late-2025) for citation generation in anterior segment ophthalmology. Among the tested versions, DeepSeek-R1 shows the most promise. Nevertheless, all models produced a substantial rate of errors, including fabricated references. Therefore, these tools cannot be relied upon for autonomous citation generation in academic or clinical writing.

The responsible path forward involves viewing AI as a preliminary aid within a workflow, dominated by human expertise and traditional, validated research methods. Future improvements in AI citation utility will depend on enhanced training on verified bibliographic corpora, integration with live databases, and the development of transparent mechanisms to flag uncertain outputs. Furthermore, given the dynamic nature of this technology, the research community should adopt frameworks for the ongoing and version-specific benchmarking of AI tools. Until such advances are realized and rigorously validated, meticulous human oversight remains the indispensable safeguard for citation integrity in medical science.

REFERENCES

1. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019;103(2):167-175. doi:10.1136/bjophthalmol-2018-313173
2. Sheng B, Chen X, Li T, et al. An overview of artificial intelligence in diabetic retinopathy and other ocular diseases. Front Public Health. 2022;10:971943. doi:10.3389/fpubh.2022.971943
3. Cheng A, Calhoun A, Reedy G. Artificial intelligence-assisted academic writing: recommendations for ethical use. Adv Simul (Lond). 2025;10(1):22. doi:10.1186/s41077-025-00350-6
4. Choudhury A, Shahsavar Y, Shamszare H. User intent to use DeepSeek for healthcare purposes and trust in large language models: multinational survey study. JMIR Hum Factors. 2025;12:e72867. doi:10.2196/72867
5. Goddard J. Hallucinations in ChatGPT: a cautionary tale for biomedical researchers. Am J Med. 2023;136(11):1059-1060. doi:10.1016/j.amjmed.2023.06.012
6. Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology–head and neck surgery. Eur Arch Otorhinolaryngol. 2024;281(4):2159-2165. doi:10.1007/s00405-023-08441-8
7. Hussain ZS, Delsoz M, Elahi M, et al. Performance of DeepSeek, Qwen 2.5 MAX, and ChatGPT in assisting diagnosis of corneal, glaucoma, and neuro-ophthalmology diseases based on clinical case reports. medRxiv [Preprint]. 2025 Mar 17. doi:10.1101/2025.03.14.25323836
8. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. J Med Internet Res. 2024;26:e53164. doi:10.2196/53164
9. Akter SN, Yu Z, Muhamed A, et al. An in-depth look at Gemini's language abilities. arXiv [Preprint]. 2023 Dec 18. arXiv:2312.11444. doi:10.48550/arXiv.2312.11444
10. Sensoy E, Citirik M. Evaluation and comparison of the knowledge levels of current artificial intelligence programs on retinal and vitreous diseases and treatment methods. J Curr Ophthalmol. 2024;36:78-81. doi:10.4103/joco.joco_192_23
11. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. Artif Intell Rev. 2023;56:3005-3054. doi:10.1007/s10462-022-10246-w
12. Noel-Storr A, Dooley G, Elliott J, et al. An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. J Clin Epidemiol. 2021;133:130-139. doi:10.1016/j.jclinepi.2021.01.006
13. Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus. 2023;15(4):e37432. doi:10.7759/cureus.37432
14. Thomo A. PubMed retrieval with RAG techniques. Stud Health Technol Inform. 2024;316:652-653. doi:10.3233/SHTI240498