Exploratory Research and Hypothesis in Medicine **2026** vol. 11(4) | e00057
DOI: 10.14218/ERHM.2026.00057

Original Article

Errors in LLM-assisted Retrieval of Medical Literature: A Comparative Study

Jenny Gao[1], Yongfeng Zhang[2], Mary L. Disis[3] and Lanjing Zhang[4,5,6*]

[1]*College of Arts and Science, New York University, New York, NY, USA;* [2]*Department of Computer Science, School of Arts & Sciences, Rutgers University, Piscataway, NJ, USA;* [3]*UW Medicine Cancer Vaccine Institute, University of Washington, Seattle, WA, USA;* [4]*Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, NJ, USA;* [5]*Department of Pathology, Princeton Medical Center, Plainsboro, NJ, USA;* [6]*Rutgers Cancer Institute, New Brunswick, NJ, USA*

Received: September 11, 2026 | **Revised:** October 22, 2026 | **Accepted:** October 28, 2026 | **Published online:** October 30, 2026

Abstract

Background and objectives: Large language model (LLM)-assisted literature retrieval may lead to bibliographic errors that have not been rigorously quantified. Therefore, we aimed to analyze errors in reference retrieval of free-version LLMs and their associated factors.

Methods: We evaluated 2,000 references retrieved by 5 LLMs (Grok, ChatGPT, Gemini, Perplexity, and DeepSeek) for 40 randomly selected original articles (10 per journal) published from January 2024 to July 2025 in *British Medical Journal, Journal of the American Medical Association, Lancet, and New England Journal of Medicine*. The primary outcomes were the score ratio, a composite score of validity of digital object identifier, PubMed ID, Google Scholar link, and relevance, with a bonus point for retrieving the index article, and the complete miss status that indicated failures on all metrics. Multilevel mixed-effects multivariable regression was used.

Results: References retrieved by LLMs failed all metrics 48.7% of the time. Grok was linked to a higher score ratio than ChatGPT (coefficient = 0.14, 95% confidence interval [CI], 0.07–0.21; $P < 0.001$), while all other platforms were associated with lower ones. Compared with ChatGPT, Grok was associated with lower odds of complete miss (better performance, odds ratio (OR) = 0.23, 95% CI, 0.11–0.46; $P < 0.001$), but DeepSeek, Gemini, and Perplexity were associated with higher odds (OR = 3.49, 95% CI, 1.87–6.50; OR = 16.86, 95% CI, 8.69–32.71; OR = 11.01, 95% CI, 5.79–20.94; respectively; $P < 0.001$ for all). Journals were linked to the score ratio, not complete miss.

Conclusions: We show that the performance of LLM-assisted literature search and retrieval in medicine is overall variable and often unreliable.

Introduction

The rapid proliferation of large language models (LLMs) has transformed numerous fields, including biomedicine and bioinformatics, where they offer potential for accelerating literature synthesis, reference retrieval, and knowledge discovery.[1,2] LLMs can assist in literature synthesis tasks such as generating relevant citations for research articles.[3,4] LLM-based ensemble strategies with retrieval-augmented generation prompts can also help literature screening.[5] However, LLMs often produce plausible but fabricated or inaccurate content, including fictitious references and incorrect digital object identifiers (DOIs).[6-9] Some scholars refer to this phenomenon as referential hallucination.[7,10-13]

High rates of referential hallucination have been documented across early LLM versions, with studies reporting fabrication in 50%–80% of generated citations in biomedical contexts.[13,14] Literature search errors have been observed not only in factual claims but also in bibliographic metadata, including DOIs and journal citations.[7,13] This observation raises concerns about reliability in high-stakes applications like medical research, where inaccurate references could propagate errors or undermine scholarly integrity.[15,16] However, there are few quantitative comparative or

Keywords: Large language models; Artificial intelligence; Literature retrieval; Bibliographic accuracy; Referential hallucination; Biomedical literature.
***Correspondence to:** Lanjing Zhang, Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers University, 164 Frelinghuysen Road, Piscataway, NJ 08854, USA. ORCID: https://orcid.org/0000-0001-5436-887X. Tel: +1-609-853-6833, Fax: +1-609-853-6841, E-mail: lanjing.zhang@rutgers.edu
How to cite this article: Gao J, Zhang Y, Disis ML, Zhang L. Errors in LLM-assisted Retrieval of Medical Literature: A Comparative Study. *Explor Res Hypothesis Med* 2026;11(4):e00057. DOI: 10.14218/ERHM.2026.00057.

© 2026 The Author(s). This article has been published under the terms of Creative Commons Attribution-Noncommercial 4.0 International License (CC BY-NC 4.0), which permits noncommercial unrestricted use, distribution, and reproduction in any medium, provided that the following statement is provided. "This article has been published in *Exploratory Research and Hypothesis in Medicine* at https://doi.org/10.14218/ERHM.2026.00057 and can also be viewed on the Journal's website at https://www.xiahepublishing.com/journal/erhm".

association analyses of the error/hallucination rates of contemporary LLMs.

Therefore, we quantified the errors of five widely used free-version LLMs (Grok, ChatGPT, Google Gemini, Perplexity AI, and DeepSeek) in searching and retrieving references relevant to original research articles published in four leading medical journals, namely *British Medical Journal* (BMJ), *Journal of the American Medical Association* (JAMA), *The Lancet,* and *The New England Journal of Medicine* (NEJM). These leading medical journals were chosen to present high-quality abstracts and scientific content in general medicine. By randomly sampling articles from January 2024 to July 2025 and rigorously validating 2,000 retrieved references using a multimetric framework, we aimed to provide an objective and reproducible assessment of errors in LLM-assisted general medical literature search and retrieval. These errors include fabricated references, real references with incorrect bibliographic data and irrelevant real references with correct bibliographic data. Our findings highlight significant variability in LLM performance and provide insights into the factors associated with these errors.

Materials and methods

Random selection of original research articles

Five original research articles published in 2024 and 5 in early 2025 (January 1 to July 15) were randomly selected from each of the four leading medical journals: BMJ, JAMA, *Lancet,* and NEJM (Fig. 1). Specific inclusion and exclusion criteria were thus not used due to the random selection process. Specifically, all original articles published in these journals in 2024 and early 2025 were assigned a number sequentially according to their PubMed indexing times, separately for each journal and publication year. For each journal, we used Google's random number generator to generate five random numbers for the 2024 cohort and another five for the 2025 cohort. Only the articles corresponding to these random numbers were included in the analysis.

Retrieval of relevant articles using LLMs

For each of the 40 original research articles, we prompted 5 commonly used LLMs (all free versions) with the article's abstract to retrieve 10 relevant references between July 15 and August 30, 2025. The LLMs were Grok (Grok-2), ChatGPT (GPT-4o), Google Gemini (2.5 Flash), Perplexity AI, and DeepSeek (V3). Each query focused on one article at a time. All 5 LLMs were accessed through free accounts without paid subscriptions and were evaluated under identical settings. The same web browser (Safari 18.6) without plug-ins or user-activated web search was applied to all 5 platforms at a given time.

The prompt was "Generate a list of 10 key references related to this abstract with the title, year published, DOI, PubMed ID, and Google Scholar URL; no description; clearly indented for each reference" with the article's abstract appended to the prompt. We then instructed the LLMs to confirm these bibliographic identifiers using the prompt "Please confirm the DOI, PubMed ID and Google Scholar link." The results confirmed by the platform were recorded and subjected to manual checking. The same prompting process was repeated in all experiments.

Manual validation of retrieved references

Four validity metrics were manually assessed using two strategies (Fig. 1).

One reviewer first searched the DOI registry (doi.org), PubMed (pubmed.gov), and Google Scholar with the retrieved reference's title and authors. If a DOI, PubMed ID or Google Scholar link was found and matched to that provided by the LLM, one point was assigned to the respective metric. If found but not matched, no points were assigned. If no entry was found, the metric was considered not applicable. For example, an engineering reference/article understandably might not be indexed in PubMed and hence had no PubMed ID. Three of the four metrics were assigned using this strategy.

In the other strategy, one reviewer compared the retrieved references with the title and reference list of the index article to assign a relevance score. If the title and first author of the retrieved reference were found in the reference list, one relevance point was assigned. If not found, no points were assigned. However, if the retrieved reference was the index article, two points were assigned to distinguish retrieval of the index article from retrieval of other relevant references.

A second reviewer independently evaluated 500 randomly selected retrieved references (100 from each platform) between December 20 and December 30, 2025. The scores were compared with the original scores. A score difference in any of the 4 metrics was considered a scoring discrepancy and was resolved by the two reviewers via discussion and fact-checking. No major LLM version changes occurred, while DeepSeek had a minor version upgrade (from V3.1 to V3.2).

Calculation of the score ratio and complete miss status

Because some retrieved references had nonapplicable metrics, we used a composite ratio and complete miss status to more objectively evaluate the retrieved references.

First, we generated the total score by summing the scores for DOI, PubMed ID, Google Scholar link and relevance score for each retrieved reference (Fig. 1). Second, we generated the score cap by counting the number of applicable validity metrics. The retrieved references with nonapplicable metrics then had a lower maximum score than those without. Finally, we divided the total score by the score cap to generate the score ratio for each retrieved reference. For example, one reference retrieved by ChatGPT for a BMJ article was indeed the index article. DOI, PubMed ID and Google Scholar link were all correct. Therefore, the reference's score cap was 4, while its total score was 5 including a bonus point for relevance. The score ratio was thus 1.25 (5 divided by 4). To highlight the retrieved references without any validated metrics, the complete miss status was assigned if the retrieved reference failed all metrics, indicating that none of the applicable validity metrics received a point.

Regression analyses and statistics

To account for the hierarchical clustering of multiple retrieved references within individual index articles and LLMs, we used mixed-effects multivariable linear and logistic regression models for the score ratio and complete miss outcomes, respectively. Index articles and clusters (article-by-LLM-platform) were

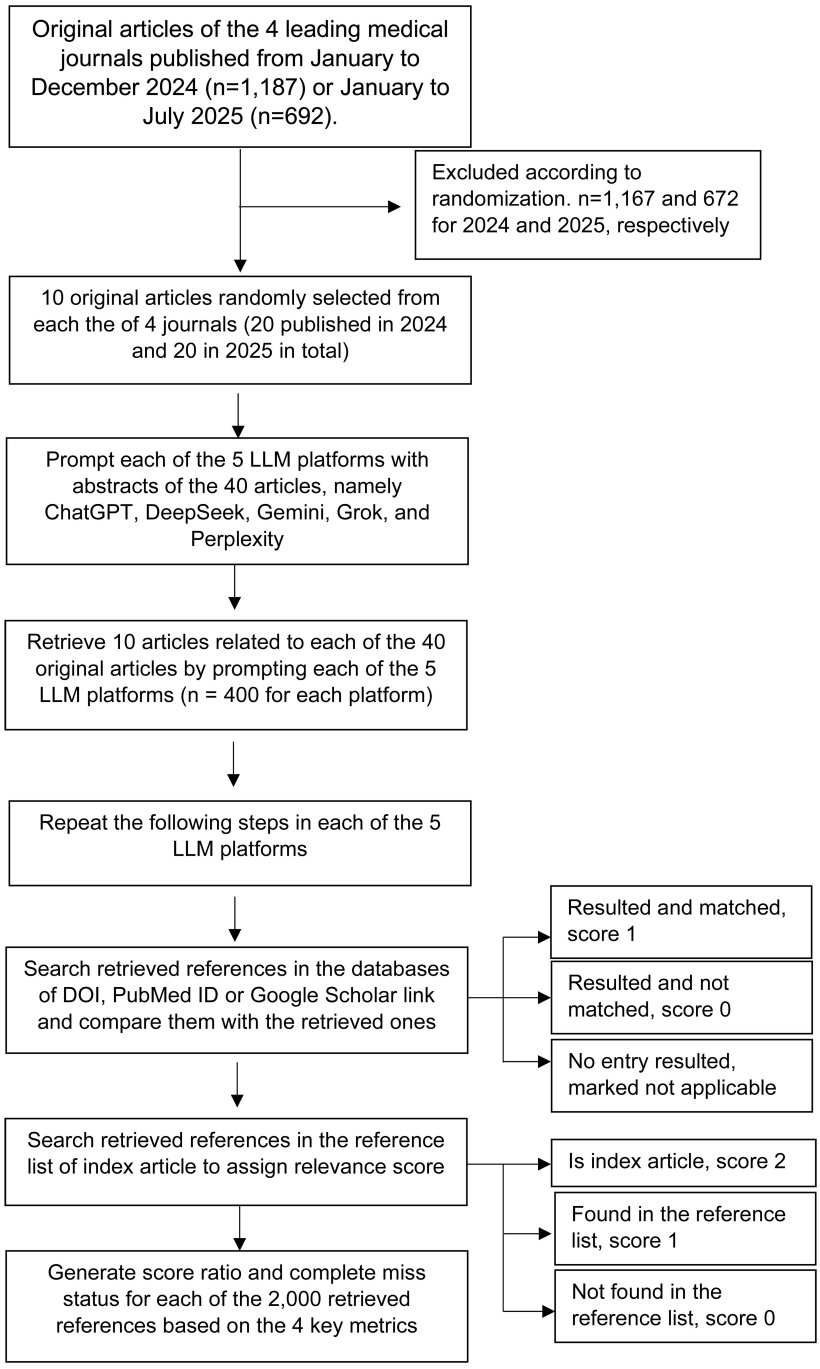


Fig. 1. Flowchart of the study. The four leading medical journals include *British Medical Journal, Journal of the American Medical Association, Lancet,* and *New England Journal of Medicine.* The four key validity metrics were digital object identifier, PubMed ID, Google Scholar link, and relevance score. DOI, digital object identifier; LLM, large language model.

modeled with random intercepts to account for baseline article-level and cluster-level heterogeneity. The fixed-effect factors of interest were journal, publication year, and LLM. These factors were evaluated as fixed predictors to quantify their specific associations if they met the inclusion criterion. We included only the factors potentially associated with either of the two outcomes in univariable regression. This liberal screening threshold of $P < 0.25$ was used according to the purposeful-selection approach described previously.[17,18] A likelihood-ratio test was then used to examine a nested random intercept for LLM within article.

The chi-square or Fisher's exact test was used to compare categorical data by metrics. We performed all statistical analyses

Table 1. Multilevel mixed-effects multivariable regression on the associations of LLM and medical journal with the score ratio

	Coefficient	95% CI	*P*
LLM			
ChatGPT	Reference		
DeepSeek	−0.21	(−0.28 to −0.14)	<0.001
Gemini	−0.31	(−0.38 to −0.24)	<0.001
Grok	0.14	(0.07 to 0.21)	<0.001
Perplexity	−0.30	(−0.36 to −0.23)	<0.001
Journal			
BMJ	Reference		
JAMA	−0.04	(−0.11 to 0.02)	0.187
Lancet	−0.06	(−0.12 to 0.00)	0.065
NEJM	−0.08	(−0.14 to −0.02)	0.013

BMJ, *British Medical Journal*; CI, confidence interval; JAMA, *Journal of the American Medical Association*; LLM, large language model; NEJM, *New England Journal of Medicine*.

and plotting with Stata (version 18; StataCorp LLC, College Station, TX, USA) and defined statistical significance using $P < 0.05$.

Results

Random selection of index articles

The 4 leading medical journals published 1,187 original articles from January 1 to December 31, 2024 and 692 from January 01 to July 15, 2025, from which we randomly selected 40 original articles (Fig. 1). The two reviewers had score differences in 9 (1.8%) of the 500 retrieved references reviewed by both. Their agreement rate was 98.2%. There were 4 score differences in Google Scholar link including 3 correct links mislabeled as incorrect and 1 incorrect link mislabeled as correct (net gain of 2 correct ones), 3 in PubMed ID and 2 in DOI.

Score ratio

The average score ratio of the 5 LLMs ($n = 400$ for each platform) was 0.29 (standard deviation, 0.35; range, 0–1.25), with a higher score ratio indicating greater accuracy in retrieving bibliographic data of medical literature. The highest and lowest score ratios were achieved by Grok (0.57) and Gemini (0.11), respectively, while the journal BMJ (0.33) appeared to have a higher score ratio than *Lancet* (0.29), NEJM (0.25) and JAMA (0.29).

Journals and LLMs were both independently associated with the score ratio in multivariable regression analyses (Table 1), while publication year was not associated with score ratio even in univariable regression. Compared with BMJ, NEJM was associated with a lower score ratio (coefficient = −0.08, 95% confidence interval [CI], −0.14 to −0.02; $P = 0.013$), but no significant differences were observed for *JAMA* or *Lancet* (coefficient = −0.04, 95% CI, −0.11 to 0.02; $P = 0.187$ for JAMA and coefficient = −0.06, 95% CI, −0.12 to 0.00; $P = 0.065$ for *Lancet*). Grok was linked to a higher score ratio than ChatGPT (coefficient = 0.14, 95% CI, 0.07 to 0.21; $P < 0.001$), while all other platforms were associated with lower score ratios (coefficient = −0.21, 95% CI, −0.28 to −0.14 for DeepSeek, coefficient = −0.31, 95% CI, −0.38 to −0.24 for Gemini, and coefficient = −0.30, 95% CI, −0.36 to −0.23 for Perplexity; $P < 0.001$ for all).

The likelihood-ratio test on the models for score ratio showed a nested random intercept for LLM within article (chi-square = 170.29, likelihood-ratio $P < 0.001$). The cluster-level intraclass correlation coefficient (ICC) was 0.21 (95% CI 0.16–0.26), while the article-level ICC was <0.01 (95% CI <0.01–0.95) and was considered trivial.

There were differences in adjusted associations between LLMs and the score ratio in most pairwise comparisons by LLMs (Fig. 2). All of our pairwise comparisons were statistically significant except Gemini versus Perplexity ($P = 0.528$).

Complete miss rate

References retrieved by LLMs failed all metrics in 48.7% of cases, as indicated by the mean complete miss rate. The lowest (best) complete miss rate was achieved by Grok (11.3%), and the highest (worst) by Gemini (78.5%). NEJM had the highest complete miss rate (56.0%), whereas rates for the other journals were similar.

Journals and LLMs were associated with complete miss rate in univariable regression, but publication year was not. Our multivariable regression analyses showed that LLMs, but not journals, were associated with complete misses (Table 2). Compared with ChatGPT, Grok was associated with a lower complete miss rate (better performance, odds ratio (OR) = 0.23, 95% CI, 0.11–0.46; $P < 0.001$), but others were associated with higher rates (worse performance, OR = 3.49, 95% CI, 1.87–6.50 for DeepSeek, OR = 16.86, 95% CI, 8.69–32.71 for Gemini, and OR = 11.01, 95% CI, 5.79–20.94 for Perplexity; $P < 0.001$ for all).

The likelihood-ratio test of the models for complete miss also showed a nested random intercept for LLM within article (chi-square = 119.97, likelihood-ratio $P < 0.001$). The cluster-level and article-level ICCs were 0.30 (95% CI, 0.23–0.38) and <0.01 (95% CI, <0.01–0.99), respectively. Given its small value and wide 95%

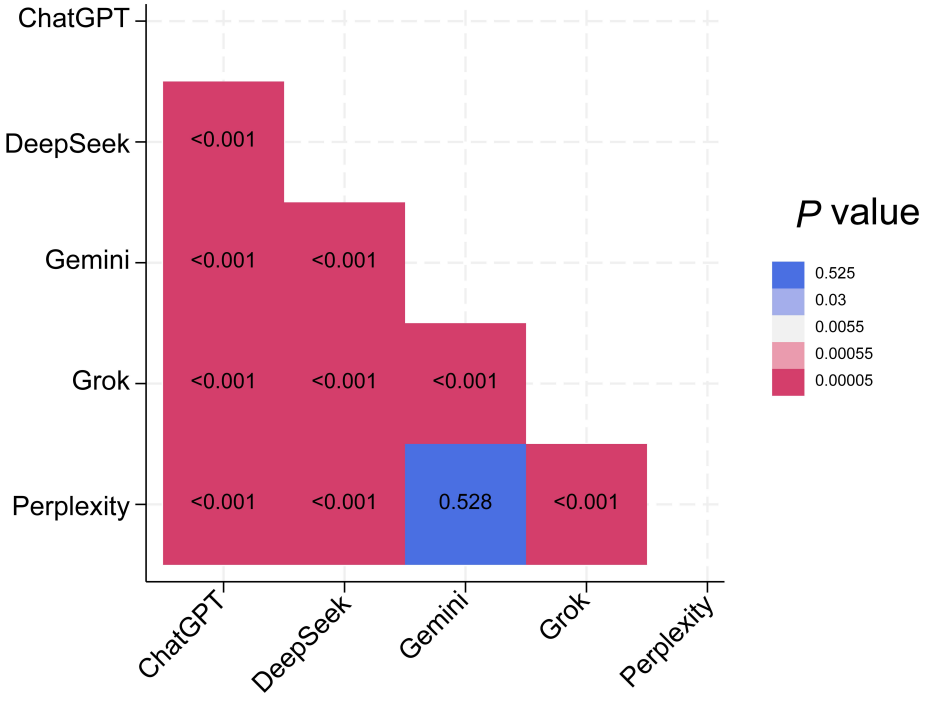


Fig. 2. Pairwise comparison of the associations between LLMs and score ratios. The associations were calculated using multivariable mixed-effects regression with journal as the covariate. LLM, large language model.

Table 2. Multilevel mixed-effects multivariable logistic regression on the associations of LLM and medical journal with complete miss

	Odds ratio	95% CI	*P*
LLM			
ChatGPT	Reference		
DeepSeek	3.49	(1.87–6.50)	<0.001
Gemini	16.86	(8.69–32.71)	<0.001
Grok	0.23	(0.11–0.46)	<0.001
Perplexity	11.01	(5.79–20.94)	<0.001
Journal			
BMJ	Reference		
JAMA	0.88	(0.48–1.60)	0.675
Lancet	0.87	(0.48–1.58)	0.657
NEJM	1.64	(0.90–3.00)	0.104

BMJ, *British Medical Journal*; CI, confidence interval; JAMA, *Journal of the American Medical Association;* LLM, large language model; NEJM, *New England Journal of Medicine.*

CI, article-level ICC should be considered trivial and characterized by substantial uncertainty.

Individual validity metrics

Across the 400 references retrieved by each LLM (100 per journal), the performance of LLMs varied significantly across all individual validity metrics according to journal or publication year (Table 3), except relevance score by publication year.

LLMs significantly differed in retrieving valid DOIs, PubMed IDs, and Google Scholar links by journal ($P < 0.001$ for all) and publication year ($P = 0.001$, 0.003 and 0.005), respectively. They also differed in relevance scores by journal ($P = 0.048$), but not

Table 3. Characteristics of LLMs' performance in retrieving references relevant to the index articles randomly selected from four leading medical journals

	ChatGPT	DeepSeek	Gemini	Grok	Perplexity	Total	*P* value
DOI							
Journal							
BMJ (*n* = 100)	62	23	3	73	9	170	<0.001
JAMA (*n* = 100)	49	13	13	72	19	166	
Lancet (*n* = 100)	35	23	30	53	22	163	
NEJM (*n* = 100)	47	11	7	55	5	125	
Year							
2024 (*n* = 200)	118	26	25	111	32	312	0.001
2025 (*n* = 200)	75	44	28	142	23	312	
PubMed ID							
Journal							
BMJ (*n* = 100)	47	13	0	50	5	115	<0.001
JAMA (*n* = 100)	25	4	4	29	12	74	
Lancet (*n* = 100)	31	10	19	34	10	104	
NEJM (*n* = 100)	26	6	4	33	8	77	
Year							
2024 (*n* = 200)	73	9	16	66	24	188	0.003
2025 (*n* = 200)	56	24	11	80	11	182	
Google Scholar link							
Journal							
BMJ (*n* = 100)	78	69	0	88	13	248	<0.001
JAMA (*n* = 100)	66	31	4	92	25	218	
Lancet (*n* = 100)	16	43	40	86	5	190	
NEJM (*n* = 100)	60	36	12	80	4	192	
Year							
2024 (*n* = 200)	115	86	39	166	16	422	0.005
2025 (*n* = 200)	105	93	17	180	31	426	
Relevance score							
Journal							
BMJ (*n* = 100)	22	16	3	46	7	94	0.048
JAMA (*n* = 100)	26	12	9	32	12	91	
Lancet (*n* = 100)	12	5	9	15	12	53	
NEJM (*n* = 100)	21	9	9	25	8	72	
Year							
2024 (*n* = 200)	41	21	20	65	21	168	0.621
2025 (*n* = 200)	40	21	10	53	18	142	

Each number represents the count of valid or relevant references for the performance metric. BMJ, *British Medical Journal*; DOI, digital object identifier; JAMA*, Journal of the American Medical Association;* LLMs, large language models*;* NEJM*, New England Journal of Medicine. P* values were calculated using Stata's Chi-square or Fisher's exact tabulation function. Because each journal-LLM or year-LLM group has an identical sample size (*n* = 100 for journal-LLM and *n* = 200 for year-LLM), the raw cell counts double as direct representations of group percentages (relative frequencies).

publication year ($P = 0.621$).

Among the 4 journals, Gemini retrieved accurate article bibliographic data least frequently for BMJ (0/100 for both PubMed ID and Google Scholar link) and JAMA (4/100 for PubMed ID and 4/100 for Google Scholar link). In contrast, Grok and ChatGPT overall retrieved accurate data more frequently than other LLMs.

Discussion

This quantitative comparative study of LLM-assisted medical literature showed modest overall performance of LLMs and significant variability in retrieval accuracy across platforms and journals. These findings align with prior reports documenting high rates of fabricated or incorrect citations in LLM outputs,[7,8,13] and extend them by demonstrating that such issues persist even when prompts explicitly request structured bibliographic metadata and data verification. We also identified associations between LLMs and overall errors in LLM-assisted literature retrieval. Future studies are needed to differentiate and analyze the errors involving fabricated references, real references with incorrect bibliographic data and irrelevant real references with correct bibliographic data.

Our findings have several implications for LLM-assisted medical literature research. First, LLM-assisted medical reference generation and retrieval should be manually verified, particularly when relying on a single platform. Reference metadata, such as DOI, PubMed ID and Google Scholar links, are often inaccurate. That is why we included these three commonly used metadata elements in the composite score, while adding the relevance score to assess the retrieved article's relevance. The bonus point assigned to the retrieved index article emphasizes the importance of retrieving the most relevant article (i.e., the index one). Variability in LLM performance, as shown here and in previous studies,[4,19] also highlights the value of using multiple LLMs at a time. This practice will not only provide more references but also offer opportunities for cross-checking.

Second, bibliographic data should always be checked even in the LLM-retrieved references with correct titles and sound relevance. Earlier studies have reported the retrieval of irrelevant or less-relevant references by LLMs.[10] Attention has therefore focused on the relevance of retrieved references. However, we show that contemporary LLMs approximate topical relevance modestly but struggle more with retrieving bibliographic data. The frequent retrieval of plausible references with incorrect bibliographic data could be misleading, offer false assurance, and reinforce the ethical and practical concerns about unsupervised use of LLM-retrieved references.[15,20]

Third, two index-article publication-year cohorts, 2024 and 2025, were used to account for possible association between publication timing (latency) and retrieval performance. However, our univariable regression analyses showed no association of publication year with score ratio or complete miss rate. Publication year was thus not included in the multivariable regression analyses and was unlikely to be associated with retrieval accuracy.

Limitations

This study has limitations. We evaluated only free versions of LLMs and focused on a simple prompt design, which may not capture the full range of model capabilities. LLM capabilities also evolve rapidly and may have reduced their errors as this study is being reviewed. Moreover, our relevance scoring, while systematic and mostly objective (e.g., citation by the index article), involved human judgment and may be subjective and biased. Citation by the index article is objective but may be too specific and restrictive, since some relevant articles may not be cited due to space limitations. However, additional manual assessment may introduce subjectivity and biases. Furthermore, only leading general medical journals were used here. Our findings thus cannot be generalized to other medical specialties without validation. They may not be directly applicable to the abstracts in middle- or lower-tier medical journals, either, due to differences in abstract quality and editorial preferences. In addition, we did not systematically examine the influence of integrated web search on free-tier LLMs, while recent studies show that web search may help improve performance of LLMs.[21,22] Besides the limited scope of this study, the key reason is that the examined LLMs differ significantly by web search availability. Perplexity and ChatGPT offer paid integration of web search. DeepSeek requires users to take manual action (toggle the button) for its free web search function, while Grok and Gemini appeared to integrate the search through X and Google for free, respectively. Our choice of leaving out user-activated web search therefore provides a desired baseline assessment of these platforms, while future work is needed. The inherent differences in LLMs' database coverage of DOI, PubMed, and Google Scholar may also affect the interpretation of individual validity metrics. Finally, only abstracts of the index articles were subject to the LLM search but may not be the best prompting strategy. However, recent studies show that abstract-based and compressed prompts lead to more accurate literature retrieval.[23-25] We believe that abstract-based prompting is more efficient and may be more accurate than the full-text-based approach. Future studies using full articles and other prompting strategies will be interesting and warranted, but are beyond the scope of this study. The findings should therefore be interpreted within the evaluated platforms, prompts, journals, and study periods.

Future directions

Explicit prompts to double-check references were used here but appeared ineffective in improving retrieval accuracy. Therefore, such prompts should be used with caution. More research is needed to better understand and improve prompting in LLM-assisted literature retrieval. One closely related question is whether abstract is more beneficial for LLM-assisted literature search and retrieval than full-text in medical fields. Another question is how to best use yet safeguard LLM-assisted literature search and retrieval in medical and other fields. It will also be interesting to compare the performances of paid versions of LLMs with those of free ones, while the paid versions may afford the users more tokens, longer uninterrupted runs and better LLMs than the free ones.

Conclusions

While LLMs hold promise as tools for biomedical literature exploration, their performance in the medical literature search and retrieval tasks evaluated here is highly variable and often unreli-

able. Until robust hallucination mitigation, verification systems, and trustworthy AI frameworks are widely implemented, LLMs should be used as assistive, not authoritative, tools in biomedical scholarship.

Acknowledgements

None.

Funding

This work was supported by the National Cancer Institute, National Institutes of Health (R37CA277812 to LZ). The funder of the study had no role in study design, data collection, analysis, interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Conflict of interest

Lanjing Zhang is a co-editor-in-chief of this journal. The authors declare no other conflict of interest, particularly with commercial LLMs.

Author contributions

Study conceptualization and design, ensuring data access, accuracy, and integrity (LZ), data collection, data analysis, and reference validation (JG, LZ), manuscript writing and revision (JG, LZ, YZ, MLD). All authors approved the final publication version.

Ethical statement

This study did not involve human participants.

Data sharing statement

The full data are available from the corresponding author upon reasonable request.

References

[1] Yang T, Xiao Y, Bao Z, Hao J, Peng J. The rise and potential opportunities of large language model agents in bioinformatics and biomedicine. Brief Bioinform 2025;26(6):bbaf601. DOI: 10.1093/bib/bbaf601, PMID: 41214870.

[2] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930–1940. DOI: 10.1038/s41591-023-02448-8, PMID: 37460753.

[3] Wang L, Li J, Zhuang B, Huang S, Fang M, Wang C, *et al*. Accuracy of Large Language Models When Answering Clinical Research Questions: Systematic Review and Network Meta-Analysis. J Med Internet Res 2025;27:e64486. DOI: 10.2196/64486, PMID: 40305085.

[4] Scherbakov D, Hubig N, Jansari V, Bakumenko A, Lenert LA. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. J Am Med Inform Assoc 2025;32(6):1071–1086. DOI: 10.1093/jamia/ocaf063, PMID: 40332983.

[5] Li Y, Du X, Wang Y, Chen X, Zhou Z, Lian J, *et al*. AI-assisted literature screening: A hybrid approach using large language models and retrieval-augmented generation. Int J Med Inform 2026;207:106205. DOI: 10.1016/j.ijmedinf.2025.106205, PMID: 41325716.

[6] Alansari A, Luqman H. Large language models hallucination: a comprehensive survey. Comput Sci Rev 2026;61:100970. DOI: 10.1016/j.cosrev.2026.100970.

[7] Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, *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, PMID: 38776130.

[8] Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. Cureus 2023;15(5):e39238. DOI: 10.7759/cureus.39238, PMID: 37337480.

[9] 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, PMID: 37182055.

[10] Kuculmez O, Usen A, Ahi ED. Referential hallucination and clinical reliability in large language models: a comparative analysis using regenerative medicine guidelines for chronic pain. Rheumatol Int 2025;45(10):240. DOI: 10.1007/s00296-025-05996-z, PMID: 41014370.

[11] King MR. An Update on AI Hallucinations: Not as Bad as You Remember or as You've Been Told. Cell Mol Bioeng 2025;18(6):543–548. DOI: 10.1007/s12195-025-00874-x, PMID: 41328307.

[12] Dumit J, Roepstorff A. AI hallucinations are a feature of LLM design, not a bug. Nature 2025;639(8053):38. DOI: 10.1038/d41586-025-00662-7, PMID: 40038472.

[13] Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, *et al*. Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study. JMIR Med Inform 2024;12:e54345. DOI: 10.2196/54345, PMID: 39083799.

[14] Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 2023;13(1):14045. DOI: 10.1038/s41598-023-41032-5, PMID: 37679503.

[15] Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened. J Med Internet Res 2023;25:e46924. DOI: 10.2196/46924, PMID: 37256685.

[16] Elali FR, Rachid LN. AI-generated research paper fabrication and plagiarism in the scientific community. Patterns (N Y) 2023;4(3):100706. DOI: 10.1016/j.patter.2023.100706, PMID: 36960451.

[17] Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. Source Code Biol Med 2008;3:17. DOI: 10.1186/1751-0473-3-17, PMID: 19087314.

[18] Bendel RB, Afifi AA. Comparison of stopping rules in forward "stepwise" regression. J Am Stat Assoc 1977;72(357):46–53. DOI: 10.1080/01621459.1977.10479905.

[19] Mostafapour M, Fortier JH, Pacheco K, Murray H, Garber G. Evaluating Literature Reviews Conducted by Humans Versus ChatGPT: Comparative Study. JMIR AI 2024;3:e56537. DOI: 10.2196/56537, PMID: 39159446.

[20] Williams AS. Fabricating citations: the policies of New Jersey public institutions of higher education. J Acad Ethics 2025;23(3):423–440. DOI: 10.1007/s10805-024-09564-1.

[21] Li X, Jin J, Dong G, Qian H, Wu Y, Wen JR, *et al*. WebThinker: empowering large reasoning models with deep research capability. In: Belgrave D, Zhang C, Lin H, Pascanu R, Koniusz P, Ghassemi M, *et al.*, editors. Advances in Neural Information Processing Systems; 2025 Dec 2-7; San Diego, CA, USA, and 2025 Nov 30-Dec 5; Mexico City, Mexico. Vol. 38. Red Hook, NY, USA: Curran Associates Inc; 2025:133164–133204. DOI: 10.52202/085713-4011.

[22] Shi X, Liu J, Liu Y, Cheng Q, Lu W. Know where to go: make LLM a relevant, responsible, and trustworthy searchers. Decis Support Syst 2025;188:114354. DOI: 10.1016/j.dss.2024.114354.

[23] Liskavets B, Ushakov M, Roy S, Klibanov M, Etemad A, Luke SK. Prompt compression with context-aware sentence encoding for fast and improved LLM inference. In: Walsh T, Shah J, Kolter Z, editors. Proceedings of the AAAI Conference on Artificial Intelligence; 2025 Feb 25-Mar 4; Philadelphia, PA, USA. Vol. 39, no. 23. Washington, DC, USA: AAAI Press; 2025:24595–24604. DOI: 10.1609/aaai.v39i23.34639.

[24] Xu Z, Li F, Chen AC, Wang X. ProCut: LLM prompt compression via attribution estimation. In: Potdar S, Rojas-Barahona L, Montella S, editors. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track; 2025 Nov 4-9; Suzhou, China. Stroudsburg, PA, USA: Association for Computational Linguistics; 2025:285-309. DOI: 10.18653/v1/2025.emnlp-industry.20.

[25] Chen Y, Zhang Y, Zhang L. Errors of LLM-assisted literature retrieval in environmental science: a comparison study of abstract versus full-text based prompts. arXiv 2026. DOI: 10.48550/arXiv.2610.05690.