

Do AI chatbots find what experts would? Effects of model, user role, and sample size on study retrieval for medical questions

Qingfang Liu

*National Institute on Drug Abuse Intramural Research Program,
National Institutes of Health,
Baltimore, MD, USA
E-mail: qingfang.liu@nih.gov*

Qiao Jin

*National Library of Medicine, National Institutes of Health,
Bethesda, MD, USA
E-mail: qiao.jin@nih.gov*

Joe D. Menke

*School of Information Sciences,
University of Illinois Urbana-Champaign,
Champaign, IL, USA
E-mail: jmenke2@illinois.edu*

Thorsten Kahnt

*National Institute on Drug Abuse Intramural Research Program,
National Institutes of Health,
Baltimore, MD, USA
E-mail: thorsten.kahnt@nih.gov*

Zhiyong Lu[†]

*National Library of Medicine, National Institutes of Health,
Bethesda, MD, USA
[†]E-mail: zhiyong.lu@nih.gov*

Large language model (LLM) chatbots are increasingly used to answer clinical questions with citations to relevant clinical studies. Prior research has largely focused on citation fabrication, leaving a gap in evaluating the quality of the retrieved studies and the factors driving their selection, particularly for newer models with stronger reasoning capabilities. In this study, we evaluated three recent, general-purpose LLM chatbots: Claude Sonnet 5, Gemini 3.1 Pro, and ChatGPT GPT-5.5. We prompted the models with clinical questions adapted from 20 review questions in Issues 6 and 7 of the 2026 Cochrane Database of Systematic Reviews, simulating patient, clinician, and evidence-synthesis researcher user roles. For each review question, we queried each of the three chatbots under each of the three user roles, with four independent repetitions per chatbot–role combination, yielding 720 responses in total ($3 \text{ chatbots} \times 3 \text{ user roles} \times 4 \text{ repetitions} \times 20 \text{ review questions}$). Each chatbot was asked to support its answers with primary clinical citations, which we then benchmarked against the included and excluded study sets of the corresponding Cochrane reviews. On average, a single chatbot response retrieved $39.2\% \pm 29.8\%$ (mean $\pm$ SD across all 720 responses) of Cochrane included studies, while citing $5.0\% \pm 9.4\%$ of Cochrane excluded studies. Recall of Cochrane included studies varied significantly by model and user role. ChatGPT achieved higher recall than Claude or Gemini ($63.1\% \pm 29.5\%$ vs. $37.0\% \pm 23.8\%$ vs. $17.3\% \pm 13.1\%$; blocked permutation test, $p = 2.0 \times 10^{-5}$). The researcher role yielded higher recall than the clinician or patient roles ($42.8\% \pm 30.8\%$ vs. $38.6\% \pm 28.9\%$ vs. $36.1\% \pm 29.3\%$; $p = 2.0 \times 10^{-5}$). Controlling for publication year, citations per year, and open-access status, sample size was the only independently significant predictor of retrieval (odds ratio 1.80 per 1-unit increase in log sample size, 95% CI 1.37–2.36, $p = 2.34 \times 10^{-5}$). These findings suggest that while LLM chatbots can retrieve some studies identified by expert reviewers, their performance varies by model and user role, and they exhibit a bias toward clinical trials with larger sample sizes. Collected LLM responses and analysis code are available at <https://github.com/QingfangLiu/llm-evidence-retrieval-bias>.

Keywords: Large Language Models; Evidence Retrieval; Retrieval Bias; Systematic Reviews; Cochrane Reviews; Evidence Synthesis; Medical Question Answering

1. Introduction

An increasing number of individuals are consulting chatbots to address medical questions, including requests to retrieve evidence from existing literature to better understand medical conditions,^{1,2} encouraged in part by strong chatbot performance on medical knowledge benchmarks.^{3,4} Yet the ability of general-purpose chatbots to identify and cite relevant clinical studies remains less well characterized,^{5–8} especially in light of prior findings that large language models can fabricate or hallucinate citations.^{9–13} At the same time, systematic reviews and meta-analyses synthesize primary studies selected by experts to produce higher-quality conclusions for medical questions.¹⁴ This raises an important question: how closely do the studies cited by chatbots align with those included in systematic reviews?

Some prior work has begun to evaluate how well general-purpose chatbots retrieve primary studies, generally by comparing chatbot-cited studies against the included-study lists of published systematic reviews. These evaluations generally report low recall, whereas precision and citation fabrication vary across systems and evaluation methods. Somer et al.⁵ evaluated six publicly accessible systems against the 14 studies in an obstetric meta-analysis; the best-performing system, Claude 3.7, identified five studies. Chelli et al.⁶ tested three LLMs across 11 systematic reviews of rotator-cuff interventions, reporting precision of 0–13.4% and hallucination rates of 28.6–91.4%. Gwon et al.⁷ found that ChatGPT and Bing AI identified only one and two benchmark randomized trials, respectively, out of 24 randomized trials. Sidhu et al.¹⁵ compared the authenticity, quality, and geographic provenance of over 1,200 references across nine contemporary chatbot configurations. Low et al.¹⁶ found that general-purpose LLMs produced very few relevant, evidence-based answers (2–10% of questions). However, many of these studies relied on models predating advanced reasoning and agentic systems, leaving it unclear how modern, more capable models behave in these contexts. Additionally, much prior work has focused on a single medical domain,^{5–7} limiting the generalizability of their conclusions.

Outside biomedicine, retrieval-benchmark work such as LitSearch¹⁷ showed that retrieval performance varies substantially with query specificity and query construction. PaperAsk evaluated GPT-4o, GPT-5, and Gemini-2.5-Flash across a range of scientific fields and found that citation retrieval failed in 48–98% of multi-reference queries.¹⁸ Gao et al. evaluated five LLMs for 40 randomly selected original articles and found that the models failed to retrieve correct reference data 47.8% of the time.¹⁹ Related work has probed adjacent failure modes rather than recall itself: whether generated citations actually support their associated claims,¹⁰ how often citations are outright fabricated or contain metadata errors,⁹ and how reference-generation performance varies by model, discipline, and publication recency across large literature-review corpora.²⁰ Other work asked chatbots to generate search strategies for bibliographic databases such as PubMed and Embase rather than to select studies directly, but this setup may not reflect how general users interact with chatbots.²¹

A further consideration is that different types of users (e.g., patients, clinicians, evidence synthesis researchers) may seek access to primary clinical studies for medical question answering,²² and whether such variation affects evidence retrieval remains an open question. Task-aligned role-play has been shown to help with reasoning benchmarks,²³ while assign-

ing the model a persona does not reliably improve accuracy and can sometimes reduce it.²⁴ Persona information more broadly shifts model predictions in ways that do not always track genuine human variation.²⁵ Prompt architecture and framing can induce systematic, non-neutral bias even when no persona is involved,²⁶ and patient question framing alone has been shown to change a model’s medical conclusions even when the underlying evidence is held fixed.²⁷ However, no prior study has tested whether a user’s self-identified role changes which primary studies a chatbot retrieves.

In this study, we evaluated three contemporary general-purpose chatbot systems (GPT-5.5, Claude Sonnet 5, and Gemini 3.1 Pro, all equipped with the ability to perform web searches) on medical questions adapted from 20 Cochrane review topics. For each topic, prompts were framed from the perspective of a patient, clinician, or evidence-synthesis researcher, and each role–topic combination was repeated four times. GPT-5.5 achieved the highest recall of Cochrane included-study sets, and evidence-synthesis researcher framing yielded higher recall than clinician or patient framing. After controlling for publication year, citation rate, and open-access status, larger sample size remained the only significant predictor of study retrieval. These findings indicate that chatbot-based evidence retrieval is incomplete, context-dependent, and biased toward larger clinical trials, with implications for patients who use chatbots for medical information, as well as for clinicians, LLM developers, medical AI researchers, and policymakers.^{28–30}

2. Methods

2.1. Review selection

We used Cochrane intervention reviews published in Issues 6 and 7 of the 2026 *Cochrane Database of Systematic Reviews*. These were the two most recent complete issues available at the time of the experiment and were published after the knowledge cutoff dates of all three chatbots (see below). The 20 eligible intervention reviews covered diverse clinical areas, evaluating pharmacological or biologic interventions ($n = 7$), procedural, surgical, or laboratory techniques ($n = 4$), bedside-care or clinical-management strategies ($n = 3$), nutritional supplementation ($n = 2$), rehabilitative or conservative physical treatments ($n = 2$), behavioral interventions ($n = 1$), and telehealth or service-delivery interventions ($n = 1$). We excluded protocols and reviews addressing diagnostic, prognostic, or other non-intervention questions.

2.2. Experiment

We evaluated three widely used, general-purpose chatbots using the reasoning-enabled model configurations listed in Table 1. These systems were selected to represent prominent consumer chatbot ecosystems. To provide external context for model capability, we additionally reference Humanity’s Last Exam (HLE), a multimodal, frontier-level academic benchmark comprising 3,000 expert-authored questions across dozens of subjects and designed to probe advanced knowledge and reasoning beyond standard benchmark suites. Each response was generated independently in a fresh anonymous web session with no prior conversational context. Data were collected from the U.S. East Coast between mid- and late July 2026.

Table 1. Chatbot systems, configurations, and benchmark performance.

Model	Knowledge cutoff	Release date	Reasoning configuration	HLE (no tools)	HLE (with tools)
Anthropic Claude Sonnet 5	January 2026	June 30, 2026	Medium effort; thinking enabled	43.2%	57.4%
Google Gemini 3.1 Pro	January 31, 2025	February 19, 2026	Extended thinking	44.4%	51.4%
OpenAI GPT-5.5	December 1, 2025	April 23, 2026	High reasoning	41.4%	52.2%

Note: Knowledge cutoff indicates the latest date through which a model’s documented training knowledge extends. HLE (Humanity’s Last Exam) is a multimodal benchmark comprising 3,000 expert-authored questions.

For each review question, we developed three prompts representing different user roles. Except for the user role stated at the beginning of each prompt, the wording was kept as consistent as possible. We also explicitly instructed the chatbots to rely on primary evidence (e.g., original research studies and clinical trials) rather than secondary evidence (e.g., systematic reviews, meta analyses, narrative reviews, committee opinions, practice guidelines, editorials, or commentaries) when answering. To account for variability in chatbot responses, each prompt was run four times. An example is provided in Table 2.

Table 2. Example of the three user-role prompt variants for one review question.

User role	Role-specific opening
Patient	I have high blood pressure, and I am considering taking medicine to help me lose weight.
Clinician	I am a clinician caring for people with high blood pressure who are considering taking medicine to help them lose weight.
Evidence-synthesis researcher	I am an evidence-synthesis researcher studying weight-loss medicines in people with high blood pressure.
Shared remainder of the prompt: I am trying to understand whether weight-loss medicines improve the health of people with high blood pressure. Find individual primary studies on this. Do not use systematic reviews, meta-analyses, narrative reviews, committee opinions, practice guidelines, or editorials or comments. List the primary studies at the end.	

2.3. Study identification

We used the studies included and excluded in the Cochrane review as the reference standard and analyzed the studies retrieved in each chatbot response. For each response, the unit of analysis was the set of uniquely identifiable primary studies referenced, rather than the number of citations or publications. This approach is consistent with conventions used in Cochrane reviews. Because the prompts did not specify a citation format, we considered a study identifiable when the information provided (e.g., a study or trial name, author and year, title, PMID or DOI, registry identifier, or the destination of a citation link) was sufficient to resolve the underlying study. A citation with conflicting metadata information was still matched to a study when the remaining information uniquely identified it; we report the metadata errors below. Unresolved citations received no matched-study credit.

3. Results

3.1. Number and composition of retrieved studies

Across 720 responses, the mean number of studies retrieved per response was 10.66, of which a mean of 61.9% were Cochrane-included studies, compared with 11.4% Cochrane-excluded and 26.7% classified as other. Both retrieval volume and study composition varied across chatbots (Figure 1). Claude retrieved a mean of 12.32 ± 6.24 studies per response, of which 52.3% were classified as included—the lowest included-study proportion among the three chatbots. Gemini retrieved the fewest studies per response (3.87 ± 1.66) but had the highest included-study proportion (69.0%). ChatGPT retrieved the most studies per response (15.80 ± 6.60), of which 64.6% were classified as included. In contrast to the chatbot-level differences, study composition was similar across user roles, although researcher-role responses retrieved more studies than patient- or clinician-role responses.

3.2. Included-study recall and recall consistency

Across 720 responses, the mean per-response recall of Cochrane-included studies was $39.2\% \pm 29.8\%$ (mean $\pm$ SD), while the mean proportion of Cochrane-excluded studies cited per response was $5.0\% \pm 9.4\%$. Mean included-study recall differed significantly by chatbot: $37.0\% \pm 23.8\%$ for Claude Sonnet 5, $17.3\% \pm 13.1\%$ for Gemini 3.1 Pro, and $63.1\% \pm 29.5\%$ for ChatGPT 5.5 (blocked permutation test, $p = 2.0 \times 10^{-5}$). Recall also differed by user role: $36.1\% \pm 29.3\%$ for patient, $38.6\% \pm 28.9\%$ for clinician, and $42.8\% \pm 30.8\%$ for researcher (blocked permutation test, $p = 2.0 \times 10^{-5}$). Recall consistency, measured as mean pairwise Jaccard similarity of retrieved Cochrane-included study sets across replicate responses, was highest for ChatGPT 5.5 (0.80), compared with Claude Sonnet 5 (0.61) and Gemini 3.1 Pro (0.56), and was similar across user roles (patient, 0.64; clinician, 0.65; researcher, 0.67). These results are summarized in Figure 2.

3.3. Study overlap by chatbot and user role

We examined which studies were cited across multiple chatbots or user-role conditions and which were cited exclusively within a single group. Cochrane-included and Cochrane-excluded

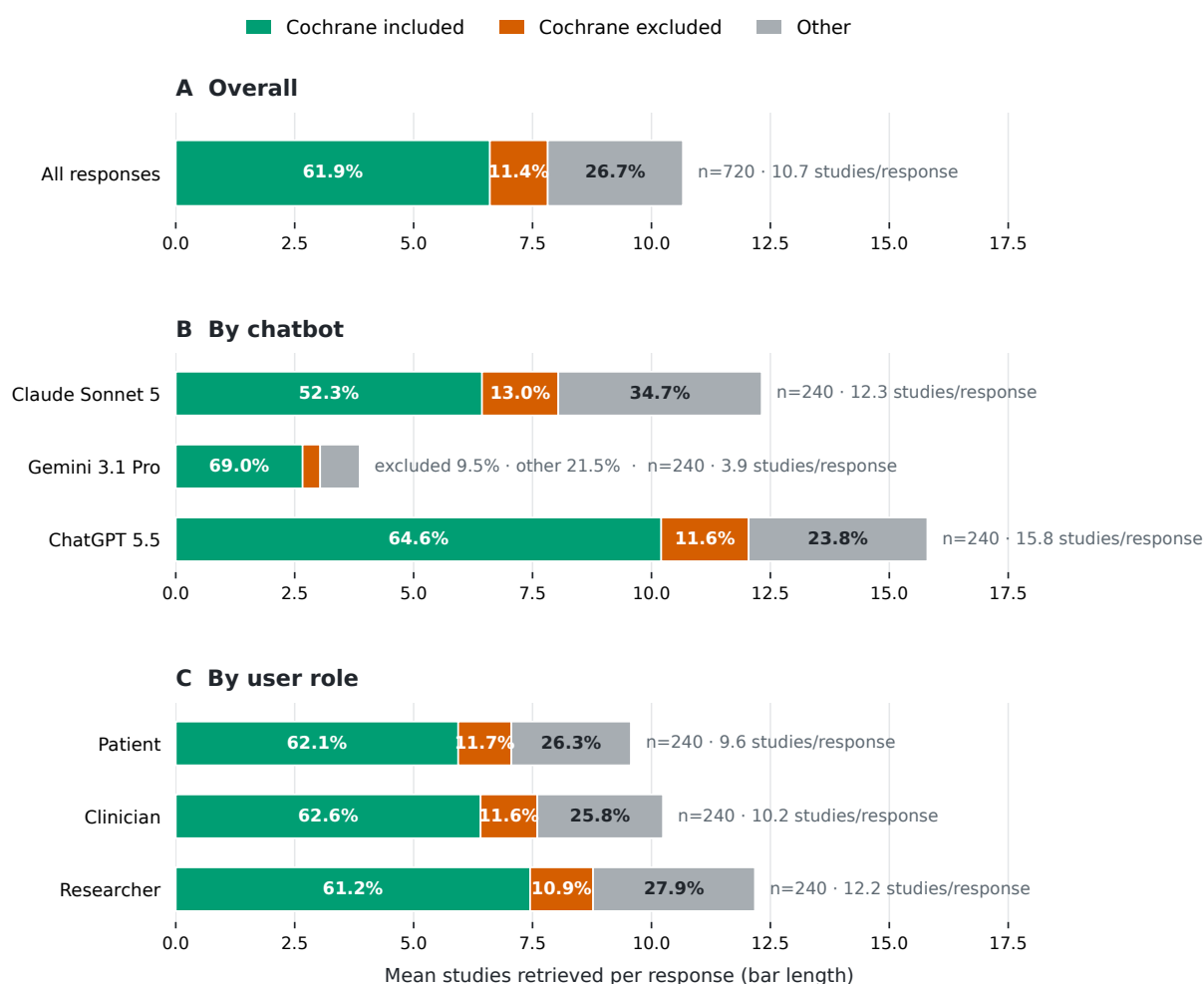


Fig. 1. **Composition and volume of studies retrieved by Cochrane status.** Stacked bars show the mean within-response proportion of distinct retrieved studies classified as Cochrane included, Cochrane excluded, or other, with bar length encoding mean studies retrieved per response on one shared scale across panels. Panel A summarizes all 720 responses; Panels B and C stratify by chatbot and user role, respectively (240 responses per group).

studies were analyzed separately (Figure 3). Of the 442 Cochrane-included studies, 328 (74.2%) were cited in at least one response. When responses were grouped by chatbot and pooled across user roles, Claude cited 236 studies (53.4%), Gemini cited 119 (26.9%), and ChatGPT cited 303 (68.6%). Among the 328 cited studies, 107 (32.6%) were cited by all three chatbots. Eighty-three studies were cited exclusively by ChatGPT, 21 exclusively by Claude, and one exclusively by Gemini; an additional 105 were cited by both Claude and ChatGPT but not by Gemini. When responses were grouped by user role and pooled across chatbots, 249 studies (56.3%) were cited under the patient role, 272 (61.5%) under the clinician role, and 314 (71.0%) under the researcher role. Of the 328 cited studies, 234 (71.3%) were cited under all three roles, whereas 4, 8, and 43 were cited exclusively under the patient, clinician, and researcher roles, respectively. Thus, the included-study sets overlapped more strongly across user roles than

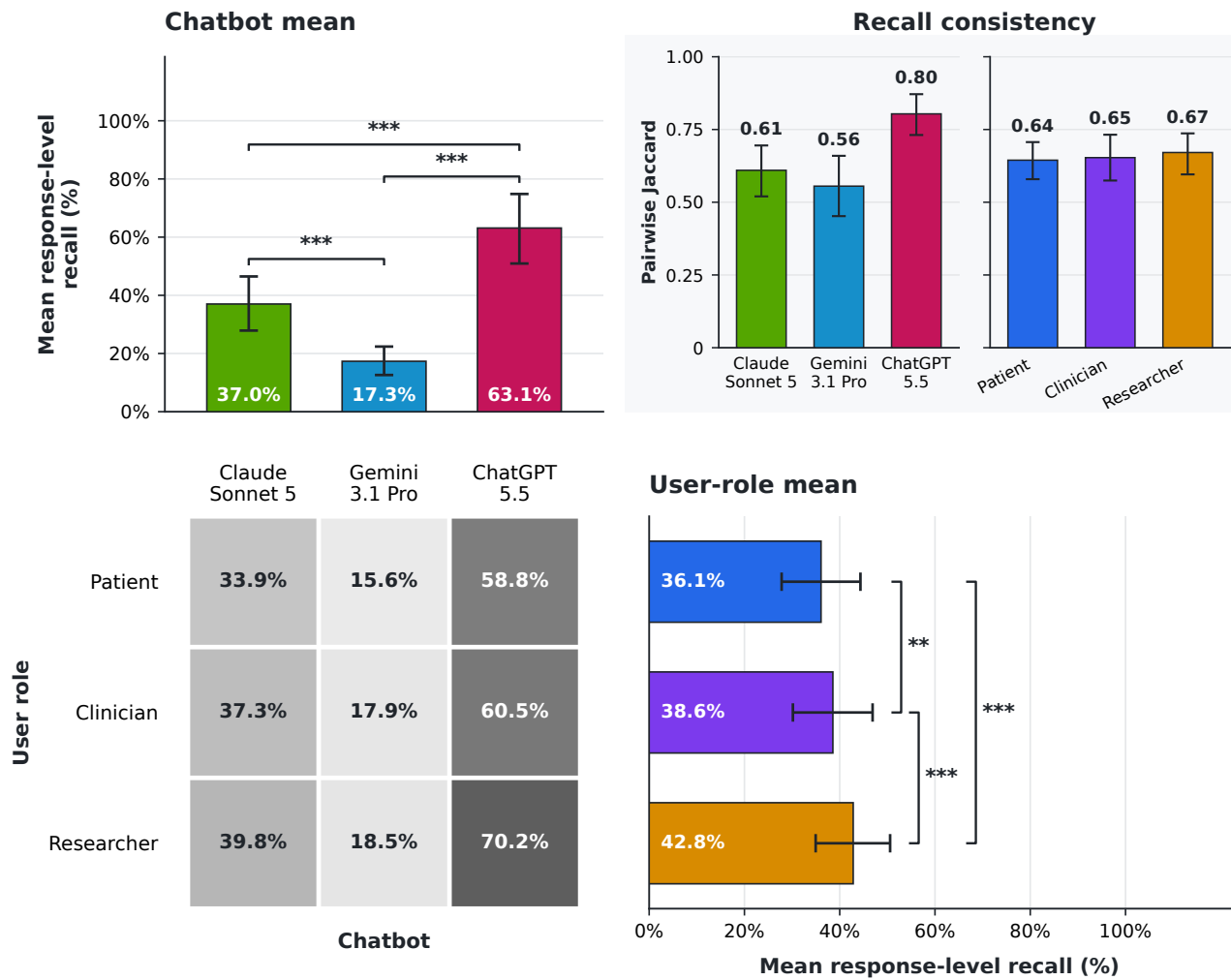


Fig. 2. **Cochrane included-study recall and recall consistency by chatbot and user role.** The heatmap shows mean response-level recall for each chatbot-by-user-role combination; marginal bars summarize recall by chatbot and by user role. The top-right panel shows recall consistency, defined as the mean pairwise Jaccard similarity of Cochrane-included study sets across replicate responses. Error bars indicate percentile 95% review-clustered bootstrap confidence intervals. Recall was calculated as the proportion of Cochrane-included studies retrieved in each response. Pairwise differences were tested using blocked permutation tests, with Holm adjustment within each marginal dimension. *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$; ns, $P \geq 0.05$.

across chatbots.

Of the 932 Cochrane-excluded studies, 143 (15.3%) were cited in at least one response, and only 16 of the 143 cited excluded studies (11.2%) were cited by all three chatbots. Across user roles, 62 of the 143 cited excluded studies (43.4%) were cited under all three roles. Thus, the excluded-study sets also overlapped more strongly across user roles than across chatbots.

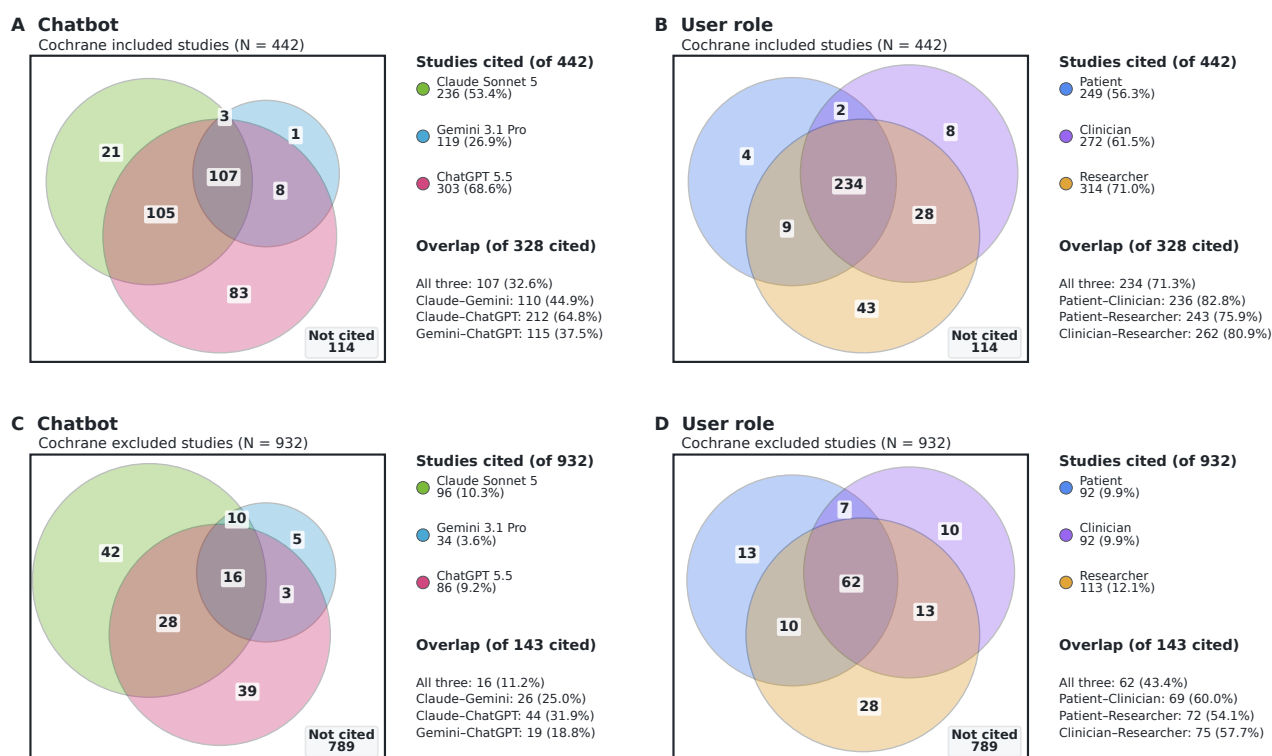


Fig. 3. **Overlap in Cochrane-included and Cochrane-excluded studies cited by chatbot and user role.** Proportional-circle Venn diagrams show citations of 442 included and 932 excluded study labels across 20 Cochrane reviews. Circle areas are proportional within each panel to the number of studies cited by each chatbot or role. Region labels report the number of studies in each mutually exclusive shared or unique citation pattern. The box enclosing each panel's circles denotes the full benchmark of included or excluded studies, and the boxed annotation at bottom right gives the number of studies not cited by any group. At right, the "Studies cited" list reports the same per-group counts and their percentage of the full benchmark, and the "Overlap" list reports the three-way and pairwise Jaccard indices (intersection over union) among the three groups.

3.4. Study characteristics associated with recall

Using the studies included in 20 Cochrane reviews as the reference standard, we compared recalled and unrecalled studies after pooling across all chatbots, user roles, and repetitions. A study was classified as recalled if it appeared in at least one response, and each Cochrane study label was counted once per review. Of 442 studies, 328 (74.2%) were recalled and 114 (25.8%) were not. Recalled studies were more recent (median publication year, 2014 vs. 2009; $p = 0.00058$), larger (median analyzed sample size, 195 vs. 97; $p < 0.001$), and more frequently cited (median citations per year, 5.50 vs. 2.33; $p < 0.001$). Open-access status did not differ significantly between recalled and unrecalled studies (56% vs. 45%; $p = 0.077$).

A multivariable logistic regression included publication year, log-transformed sample size, log-transformed citations per year, and open-access status. Among 330 studies with complete data, only sample size was significantly associated with recall (adjusted OR, 1.80 per unit increase in log sample size; 95% CI, 1.37–2.36; $p < 0.001$). Publication year, citations per year,

and open-access status were not significantly associated with recall.

Table 3. Multivariable logistic regression of study recall.

Predictor	$\hat{\beta}$	Clust. SE	Clust. p	Naive p	OR	95% CI	VIF
Intercept	−34.9314	35.1336	0.320	0.194	—	—	—
Publication year	0.0164	0.0176	0.350	0.219	1.017	(0.98, 1.05)	1.30
log(sample size)	0.5880	0.1390	< 0.001***	< 0.001	1.800	(1.37, 2.36)	1.19
log(citations/year + 0.01)	0.2993	0.1574	0.057 [†]	0.010	1.349	(0.99, 1.84)	1.54
Open access	−0.1043	0.2598	0.688	0.740	0.901	(0.54, 1.50)	1.20

Note: The model included 330 of the 442 Cochrane-included studies (74.7%) with complete data for all predictors. Standard errors and primary p -values were clustered by review (20 review clusters); naive p -values ignore clustering. OR = odds ratio; CI = confidence interval; VIF = variance inflation factor. Odds ratios and confidence intervals are not reported for the intercept. McFadden pseudo- R^2 = 0.1340; log-likelihood = −146.64; likelihood-ratio test against the null model, p < 0.001. *** p < 0.001; [†] p < 0.10.

3.5. Characterizing Cochrane-excluded studies cited by chatbots

Given that the mean proportion of Cochrane-excluded studies cited per response was $5.0\% \pm 9.4\%$, we characterized these studies by classifying the 142 unique studies into 11 categories using the exclusion reasons reported by Cochrane review authors (Table 4). All 142 studies were assigned to at least one category; 29 studies (20.4%) had multiple exclusion reasons, so the categories were not mutually exclusive. The most common reasons for exclusion concerned study design, intervention, comparator, and population. This pattern is consistent with the structure of Cochrane eligibility criteria, which are typically defined according to the population, intervention, comparator, and eligible study designs specified for each review.³¹ Because our prompts were broader than the reviews’ full eligibility criteria, some studies could be relevant to the chatbot questions yet ineligible for the more narrowly defined reviews.³²

3.6. Study metadata errors

We define *study metadata error* as a case in which a chatbot identifies a real study or study-linked publication but reports conflicting bibliographic details for it, most often the wrong lead author, year, journal, or page range. Across the 20 reviews, 232 of 7,676 matched response-study citation rows (3.0%) carried this flag. The rate differed by model: Claude flagged 5.9% of its matched citations (175/2,956), ChatGPT flagged 1.5% (55/3,792), and Gemini flagged 0.2% (2/928). By user role, the difference was smaller: patient 2.7%, clinician 3.1%, and researcher 3.2%. Beyond these metadata errors, we identified 23 additional citations that could not be resolved to any real publication, all of which came from Claude’s responses.

Table 4. Exclusion categories among 142 Cochrane-excluded studies.

Category	n/142 (%)	Definition
Study design	37 (26.1)	Nonrandomized, observational, single-arm, pseudo-randomized, post-hoc, or otherwise ineligible designs
Intervention	34 (23.9)	Ineligible intervention identity, composition, route, dose, timing, or co-intervention
Comparator	28 (19.7)	Absent or ineligible control, or an ineligible comparison between intervention variants
Population	25 (17.6)	Ineligible participants or mixed populations without separable results for the eligible subgroup
Insufficient or unavailable data	15 (10.6)	Eligibility or eligible effects could not be established, separated, or obtained
Timing or follow-up	11 (7.7)	Ineligible treatment timing, study duration, or follow-up duration
Publication or trial status	8 (5.6)	Protocols, registrations without eligible results, unpublished, withdrawn, or retracted trials, or duplicate or secondary reports
Setting	6 (4.2)	Ineligible care or recruitment setting
Unspecified eligibility	5 (3.5)	Inclusion criteria were not met, but no specific criterion was identified
Outcome	2 (1.4)	Required outcomes were absent or ineligible
Unit of analysis	2 (1.4)	Ineligible allocation or analysis unit, or non-independent observations

4. Discussion

4.1. *Study findings*

Our findings show that chatbot retrieval was incomplete and selective. A single response retrieved an average of 39.2% of Cochrane-included studies, whereas pooling all responses identified 74.2% at least once. Recall varied more by chatbot (63.1% for ChatGPT, 37.0% for Claude, and 17.3% for Gemini) than by user role (36.1–42.8%). Larger sample size was the only independent predictor of retrieval (OR = 1.80, 95% CI 1.37–2.36).

4.2. *Recall of primary studies*

Recall in this study was generally higher than that reported in several earlier evaluations of general-purpose chatbots: mean per-response recall was 39.2% overall and 63.1% for ChatGPT, compared with 0%–13.7% in Chelli et al.,⁶ only one and two of 24 benchmark trials in Gwon et al.,⁷ and 35.7% for the best-performing model in Somer et al..⁵ This difference may reflect the use of newer LLMs with stronger reasoning and agentic capabilities, an explicit prompt requesting a list of primary studies, and a user role closely matched to the medical question.

We observed differences in recall across different chatbot models, and this appears to be largely driven by citation volume. ChatGPT cited a mean of 15.80 studies per response, ap-

proximately four times that of Gemini’s 3.87. The official release materials for all three models (GPT-5.5, Claude Sonnet 5, and Gemini 3.1 Pro) emphasize advanced reasoning and agentic capabilities, and their performance on Humanity’s Last Exam was broadly comparable, especially under no-tool evaluation conditions. Despite these similarities in general reasoning performance, the models exhibited substantially different study-retrieval behaviors when answering medical questions. These discrepancies raise the question of whether they stem from differences in training data, retrieval architecture, or reasoning architecture, though the proprietary nature of these models makes this difficult to determine.

4.3. *Factors affecting retrieval*

Our study found that, when adjusting for publication year, citations per year, and open-access status, study reference retrieval was associated only with larger trial size, extending prior evidence that LLM-generated references favor highly cited publications.^{20,33} Citation rate remained positively associated with retrieval, though its clustered confidence interval included the null ($p = 0.057$). Several mechanisms may explain the preferential retrieval of larger studies. Search or retrieval-augmented generation pipelines may rank prominent “landmark” trials more highly, while larger trials may also be more strongly represented in model training data or other web-accessible sources. Chatbots may further select larger trials because sample size is interpreted as a signal of stronger or more reliable evidence. Alternatively, sample size may proxy for unmeasured characteristics such as journal visibility, trial registration, and broader scientific prominence. Because the search, retrieval, and reranking processes of these consumer chatbot systems are proprietary and opaque, the current study cannot distinguish those possible mechanisms.

4.4. *Study fabrication and metadata errors*

Although hallucination is widely recognized as a critical issue in LLM-generated citations, in practice the term functions as an umbrella covering multiple distinct phenomena:^{9,34,35} citing a completely non-existent publication, producing a citation with conflicting metadata, and the more subtle case of misattributing study content to an otherwise correct source. Here we report two categories of error: (1) metadata errors that can nonetheless be resolved to a real publication, and (2) errors that cannot be resolved to any real publication. In this experiment, the latter case was relatively rare, likely due to the strong reasoning and agentic capabilities of the models evaluated. Citation errors nonetheless persist, and Claude had a higher error rate than ChatGPT.

4.5. *Implications*

Our study has several important implications. First, patients and clinicians should recognize that a chatbot response with citations reflects a model-specific and role-conditioned subset of the literature that may vary across repeated queries, rather than a comprehensive evidence summary. Second, systematic reviewers and evidence-synthesis researchers may find that general-purpose chatbots identify some relevant studies, but these tools should not replace reproducible database searches, formal eligibility screening, or citation verification.^{36,37}

Third, AI researchers and developers working to automate evidence synthesis should make efforts to improve how systems interpret and apply population, intervention, comparator, outcome, and study-design (PICO-S) criteria. Accurate application of these criteria is essential for distinguishing the retrieval of high-quality evidence from the comprehensive retrieval of all relevant evidence.^{38,39}

4.6. *Limitations*

Our study has several limitations. First, the questions used to query the models were derived from only 20 reviews published in two recent issues of the Cochrane Reviews, limiting the breadth and generalizability of the findings and making this an exploratory test set rather than a comprehensive benchmark. Second, the study may not fully capture real-world patterns of use. Prompts were kept nearly identical across user roles to isolate the effect of this variable, but patients may express their evidence needs in more varied, less structured ways in practice.^{40,41} Models were also instructed to avoid secondary evidence to facilitate comparison with the Cochrane reference standard, but in real life, users are unlikely to add such constraints. Third, although we selected Cochrane reviews published after each model's knowledge cutoff to minimize potential training data exposure, the models may still have been exposed to earlier review versions or other secondary sources. Finally, although all models were instructed to avoid secondary evidence, model-specific differences in adherence to explicit constraints may have confounded the recall comparison.⁴²

4.7. *Future work*

Future work should evaluate citations that fall outside the sets of included and excluded studies reported in the reference Cochrane reviews. Based on the present analysis, we cannot determine whether these citations represent irrelevant studies retrieved in error, newly published eligible studies, or other forms of related evidence.⁴³ Assessing their relevance will require independent adjudication and, in some cases, clinical expertise. Because our analysis focused on study retrieval, future research should also examine whether chatbot-generated conclusions accurately reflect the underlying evidence,^{10,12} appropriately balance benefits and harms, and communicate uncertainty. Such evaluations will likely require topic-specific expertise, particularly when the evidence is heterogeneous or clinically complex. Finally, future work should evaluate agentic search systems, including Deep Research in ChatGPT and Gemini and domain-specific agentic RAG systems, to determine whether they can retrieve additional studies included in the reference Cochrane reviews that were missed by the chatbot systems evaluated in the current work.⁴⁴⁻⁴⁷

5. Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered works of the US Government. Q.J. was also supported by the NIH Pathway to Independence Award K99LM014903. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the US Department of Health and Human Services.

References

1. B. Costa-Gomes, P. Tolmachev, E. Taysom, V. Sounderajah, H. Richardson, P. Schoenegger, X. Liu, M. M. Nour, S. Spielman, S. F. Way, Y. Shah, M. Bhaskar, H. Nori, C. Kelly, P. Hames, B. Gross, M. Suleyman and D. King, Public use of a generalist LLM chatbot for health queries, *Nature Health* **1**, 689 (2026).
2. R. B. Shpiner, Why patients ask the chatbot first, *Annals of Internal Medicine* (2026).
3. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Sementurs, A. Karthikesalingam and V. Natarajan, Large language models encode clinical knowledge, *Nature* **620**, 172 (2023).
4. Q. Jin, N. Wan, R. Leaman, S. Tian, Z. Wang, Y. Yang, Z. Wang, G. Xiong, P.-T. Lai, Q. Zhu, B. Hou, M. Sarfo-Gyamfi, G. Zhang, A. Gilson, B. Bhasuran, Z. He, A. Zhang, J. Sun, C. Weng, R. M. Summers, Q. Chen, Y. Peng and Z. Lu, Tutorial: Guidance on the use of large language models for medical research, *Nature Protocols* (2026).
5. S. Somer, M. Vardy, J. Eisenberger, M. Shapira, S. Mehler and M. Safrai, Evaluation of six large language models for study identification in an obstetric systematic review, *Minerva Obstetrics and Gynecology* **78** (2026).
6. M. Chelli, J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, J.-L. Raynier, G. Clowez, P. Boileau and C. Ruetsch-Chelli, Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis, *Journal of Medical Internet Research* **26**, p. e53164 (2024).
7. Y. N. Gwon, J. H. Kim, H. S. Chung, E. J. Jung, J. Chun, S. Lee and S. R. Shim, The use of generative AI for scientific literature searches for systematic reviews: ChatGPT and Microsoft Bing AI performance evaluation, *JMIR Medical Informatics* **12**, p. e51187 (2024).
8. A. Sarker, R. Zhang, Y. Wang, Y. Xiao, S. Das, D. Schutte, D. Oniani, Q. Xie and H. Xu, Natural language processing for digital health in the era of large language models, *Yearbook of Medical Informatics* **33**, 229 (2024).
9. W. H. Walters and E. I. Wilder, Fabrication and errors in the bibliographic citations generated by ChatGPT, *Scientific Reports* **13**, p. 14045 (2023).
10. K. Wu, E. Wu, K. Wei, A. Zhang, A. Casasola, T. Nguyen, S. Riantawan, P. Shi, D. Ho and J. Zou, An automated framework for assessing how well LLMs cite relevant medical references, *Nature Communications* **16**, p. 3615 (2025).
11. X. Wang, M. Tan, Q. Jin, G. Xiong, Y. Hu, A. Zhang, Z. Lu and M. Zhang, MedCite: Can language models generate verifiable text for medicine?, in *Findings of the Association for Computational Linguistics: ACL 2025*, eds. W. Che, J. Nabende, E. Shutova and M. T. Pilehvar (Association for Computational Linguistics, Vienna, Austria, July 2025).
12. Q. Jin, Y. Fang, L. He, Y. Yang, G. Xiong, Z. Wang, N. Wan, J. Chan, D. C. Comeau, R. Leaman *et al.*, Med-v1: Small language models for zero-shot and scalable biomedical evidence attribution, *arXiv preprint arXiv:2603.05308* (2026).
13. M. Topaz, N. Roguin, P. Gupta, Z. Zhang and L.-M. Peltonen, Fabricated citations: an audit across 2· 5 million biomedical papers, *The Lancet* **407**, 1779 (2026).
14. C. Lefebvre, J. Glanville, S. Briscoe, R. Featherstone, A. Littlewood, M.-I. Metzendorf, A. Noel-Storr, R. Paynter, T. Rader, J. Thomas and L. S. Wieland, Chapter 4: Searching for and selecting studies, in *Cochrane Handbook for Systematic Reviews of Interventions*, eds. J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page and V. A. Welch (Cochrane, 2025) Version 6.5.1; last updated March 2025.

15. R. S. Sidhu, A. Selvamogan and A. Boddy, Trust, truth and transparency: Analysing the references underpinning AI-generated surgical information, *The Annals of The Royal College of Surgeons of England* (2026).
16. Y. S. Low, M. L. Jackson, R. J. Hyde, R. E. Brown, N. M. Sanghavi, J. D. Baldwin, C. W. Pike, J. Muralidharan, G. Hui, N. Alexander, H. Hassan, R. V. Nene, M. Pike, C. J. Pokrzywa, S. Vedak, A. P. Yan, D.-h. Yao, A. R. Zipursky, C. Dinh, P. Ballentine, D. C. Derieg, V. Polony, R. N. Chawdry, J. Davies, B. B. Hyde, N. H. Shah and S. Gombor, Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems, *DIGITAL HEALTH* **11**, p. 20552076251348850 (2025).
17. A. Ajith, M. Xia, A. Chevalier, T. Goyal, D. Chen and T. Gao, LitSearch: A retrieval benchmark for scientific literature search, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, 2024).
18. Y. Wu, X. Liu, Y. Feng, J. Ding and X. Ma, Paperask: A benchmark for reliability evaluation of llms in paper search and reading (2025).
19. J. Gao, Y. Zhang, M. L. Disis and L. Zhang, Errors in ai-assisted retrieval of medical literature: A comparative study, *arXiv preprint arXiv:2603.22344* (2026).
20. X. Tang, X. Duan and Z. Cai, Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition, in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, 2025).
21. W. S. W. Tam, N. Tung, S. X. Lee, G. Štiglic, T. Huynh and A. Tang, Development and evaluation of a generative AI chatbot for database searching in systematic review, *Journal of Nursing Scholarship* **58**, p. e70076 (2026).
22. M. C. Easterlin, C. T. Berdahl, S. Rabizadeh, B. Spiegel, L. Agoratus, C. Hoover and R. Dudovitz, Child and parent perspectives on the acceptability of virtual reality to mitigate medical trauma in an infusion center, *Maternal and Child Health Journal* **24**, 986 (2020).
23. A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang and X. Dong, Better zero-shot reasoning with role-play prompting, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, (Association for Computational Linguistics, 2024).
24. M. Zheng, J. Pei, L. Logeswaran, M. Lee and D. Jurgens, When “a helpful assistant” is not really helpful: Personas in system prompts do not improve performances of large language models, in *Findings of the Association for Computational Linguistics: EMNLP 2024*, (Association for Computational Linguistics, 2024).
25. T. Hu and N. Collier, Quantifying the persona effect in LLM simulations, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Association for Computational Linguistics, 2024).
26. M. Brucks and O. Toubia, Prompt architecture induces methodological artifacts in large language models, *PLOS One* **20**, p. e0319159 (2025).
27. H. S. Yun, G. Kapoor, M. Mackert, R. Kouzy, W. Xu, J. J. Li and B. C. Wallace, This treatment works, right? evaluating LLM sensitivity to patient question framing in medical QA (2026).
28. J. G. Meyer, R. J. Urbanowicz, P. C. Martin, K. O'Connor, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. J. Won, G. Gonzalez-Hernandez *et al.*, Chatgpt and large language models in academia: opportunities and challenges, *BioData mining* **16**, p. 20 (2023).
29. D. Weissenbacher, M. Shabbir, I. M. Campbell, C. T. Berdahl and G. Gonzalez-Hernandez, Enhancing medical knowledge in large language models via supervised continued pretraining on clinical notes, *medRxiv*, 2026 (2026).
30. S. S. Sahoo, J. M. Plasek, H. Xu, Ö. Uzuner, T. Cohen, M. Yetisgen, H. Liu, S. Meystre and Y. Wang, Large language models for biomedicine: foundations, opportunities, challenges, and

best practices, *Journal of the American Medical Informatics Association* **31**, 2114 (2024).

31. J. E. McKenzie, S. E. Brennan, R. E. Ryan, H. J. Thomson, R. V. Johnston and J. Thomas, Defining the criteria for including studies and how they will be grouped for the synthesis, *Cochrane handbook for systematic reviews of interventions*, **33** (2019).
32. T. Edinger and A. M. Cohen, A large-scale analysis of the reasons given for excluding articles that are retrieved by literature search during systematic review, *AMIA Annual Symposium Proceedings* **2013**, 379 (2013).
33. A. Algaba, C. Mazijn, V. Holst, F. Tori, S. Wenmackers and V. Ginis, Large language models reflect human citation patterns with a heightened citation bias, in *Findings of the Association for Computational Linguistics: NAACL 2025*, (Association for Computational Linguistics, 2025).
34. M. Dassen, R. Kotula, K. Murray, A. Yates, D. Lawrie, E. Kayi, J. Mayfield and K. Duh, Factum: Mechanistic detection of citation hallucination in long-form rag (2026).
35. M. Bhattacharyya, V. M. Miller, D. Bhattacharyya, L. E. Miller and V. Miller, High rates of fabricated and inaccurate references in chatgpt-generated medical content, *Cureus* **15** (2023).
36. J. Clark, B. L. Barton, L. Albarqouni, O. Byambasuren, T. Jowsey, J. W. Keogh, T. Liang, C. Moro, H. M. O'Neill and M. A. Jones, Generative artificial intelligence use in evidence synthesis: A systematic review, *Research Synthesis Methods* **16**, 601 (2025).
37. H. Chen, Z. Jiang, X. Liu, C. C. Xue, S. M. E. Yew, B. Sheng, Y.-F. Zheng, X. Wang, Y. Wu, S. Sivaprasad *et al.*, Can large language models fully automate or partially assist paper selection in systematic reviews?, *British Journal of Ophthalmology* **109**, 962 (2025).
38. Y. Li, S. Datta, M. Rastegar-Mojarad, K. Lee, H. Paek, J. Glasgow, C. Liston, L. He, X. Wang and Y. Xu, Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation, *Journal of the American Medical Informatics Association* **32**, 616 (2025).
39. S. K. Vallamchetla, O. Abdelkader, A. Elnaggar, D. Ramadan, M. M. I. Shourav, I. B. Riaz and M. P. Lin, Do it faster with picos: Generative ai-assisted systematic review screening, *Journal of biomedical informatics* **168**, p. 104860 (2025).
40. K. Roberts and D. Demner-Fushman, Interactive use of online health resources: a comparison of consumer and professional questions, *Journal of the American Medical Informatics Association* **23**, 802 (2016).
41. Q. Zeng, S. Kogan, N. Ash, R. A. Greenes and A. A. Boxwala, Characteristics of consumer terminology for health information retrieval, *Methods of information in medicine* **41**, 289 (2002).
42. Y. Jiang, Y. Wang, X. Zeng, W. Zhong, L. Li, F. Mi, L. Shang, X. Jiang, Q. Liu and W. Wang, FollowBench: A multi-level fine-grained constraints following benchmark for large language models, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. L.-W. Ku, A. Martins and V. Srikumar (Association for Computational Linguistics, Bangkok, Thailand, August 2024).
43. P. Garner, S. Hopewell, J. Chandler, H. MacLehose, E. A. Akl, J. Beyene, S. Chang, R. Churchill, K. Dearness, G. Guyatt *et al.*, When and how to update systematic reviews: consensus and checklist, *bmj* **354** (2016).
44. Z. Wang, Z. Chen, Z. Yang, X. Wang, Q. Jin, Y. Peng, Z. Lu and J. Sun, Empowering biomedical evidence exploration and synthesis with deep knowledge graph research, *Nature Machine Intelligence*, 1 (2026).
45. Z. Wang, C.-H. Wei, J. Chan, R. Leaman, C.-P. Day, C. Wu, M. A. Knepper, A. S. Farias, J. Rincon-Torroella, H. Slika, B. Tyler, R. H.-T. Nguyen, A. Indurkar, M. Hébert, S. Tian, L. He, N. Naffakh, A. Aseem, N. Wan, E. Y. Chew, T. D. L. Keenan and Z. Lu, Deeper-med: Advancing deep evidence-based research in medicine through agentic ai (2026).
46. Y. Xi, J. Lin, Y. Xiao, Z. Zhou, R. Shan, T. Gao, J. Zhu, W. Liu, Y. Yu and W. Zhang, A survey of large language model-based search agents, in *Proceedings of the 64th Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), eds. M. Liakata, V. P. Moreira, J. Zhang and D. Jurgens (Association for Computational Linguistics, San Diego, California, United States, July 2026).

47. J. Deng, J. Huang, Z. H. Wong, H. Liang, Q. Xu, B. Cui and W. Zhang, Data-centric perspectives on agentic retrieval-augmented generation: A survey, in *Findings of the Association for Computational Linguistics: ACL 2026*, eds. M. Liakata, V. P. Moreira, J. Zhang and D. Jurgens (Association for Computational Linguistics, San Diego, California, United States, July 2026).