LETTER

Accelerating Medical Systematic Reviews with Large Language Models: Performance, Consistency, and Potential Issues

Bo Hu[1], Ziqian Xia[2], Liuna Geng[1]

1 School of Social and Behavioral Sciences, Nanjing University, Nanjing, China
2 Doerr School of Sustainability, Stanford University, Stanford, USA

Correspondence
Liuna Geng, Email: gengliuna@nju.edu.cn.

This manuscript has been accepted by the *Journal of Evidence-Based Medicine* and can be accessed at the following link: https://doi.org/10.1111/jebm.70207.

Well-conducted systematic reviews are central to evidence-based medicine and often inform clinical guidelines and health policy. By synthesizing evidence using explicit and reproducible methods, they can minimize bias. Their production, however, is increasingly constrained by the exponential growth of biomedical literature and by severe resource limitations. A single systematic review requires, on average, 67 weeks from protocol registration to publication, together with substantial human labor to screen thousands of citations and extract granular data.[1] Traditional machine-learning and natural-language-processing tools offer partial solutions but face adoption barriers: they typically require extensive, domain-specific training data and are built for single tasks, lacking the flexibility to adapt to new topics without retraining.

Recent advances in large language models (LLMs) present an opportunity to address these bottlenecks. Unlike their predecessors, LLMs draw on knowledge pre-trained from vast text corpora to perform complex cognitive tasks through natural-language prompting, potentially streamlining the review workflow without task-specific training. [2] Yet the reliability of LLMs across the rigorous, transparent, and reproducible stages of evidence synthesis remains debated: the probabilistic nature of generative artificial intelligence (AI) raises concerns about consistency and accuracy in a field where precision is paramount. To examine this, we conducted a systematic review of empirical studies that quantitatively evaluated LLMs across five core review stages, focusing on performance, agreement with human reviewers, run-to-run reproducibility, and efficiency. Several reviews have already mapped LLM use in evidence synthesis [3–5]. The present review extends this literature by updating the evidence through 1 July 2026, reporting study-level primary and record-level sensitivity summaries with full observed ranges, separating agreement with human reviewers from run-to-run reproducibility, and examining efficiency and prompt-strategy subgroups across five core stages.

The review was not prospectively registered, and the draft protocol, together with an amendment log, is provided in the Supplementary Materials. We searched PubMed, Web of Science, and Scopus from 1 January 2021 through 25 November 2024, re-ran the full search through 5 March 2026, and conducted a further update on 1 July 2026. The 2021 lower bound was selected conservatively to capture early applications emerging with

general-purpose foundation models; it should not be interpreted as the point of widespread uptake, which accelerated after ChatGPT became publicly available in November 2022. [6,7] The strategy combined three concept blocks: (1) LLM-related terms; (2) evidence-synthesis study-design terms; and (3) terms for the five eligible review stages (literature search, title-and-abstract screening, full-text screening, data extraction, and quality assessment). For the LLM block we deliberately retained the generic terms "large language model(s)" and "LLM" alongside an expanded list of named model families (including GPT, Claude, LLaMA, Gemini, Qwen, DeepSeek, Mistral, Grok, GLM, and Phi, among others), so that studies of models not individually named could still be retrieved through the generic terms; the corrected and expanded strategy for each database is reported in the Supplementary Materials. We nonetheless acknowledge that no search can wholly exclude the risk of missing studies of newly released models. Inclusion was restricted to English-language reports. We included empirical primary studies reporting an original quantitative evaluation of publicly accessible LLMs in medical evidence synthesis. Secondary reviews were excluded as units of inclusion; the reference lists of relevant reviews were hand-searched, and any potentially eligible primary studies identified through this process were assessed using the same eligibility criteria. Two reviewers independently screened titles and abstracts (Cohen's $\kappa = 0.85$) and then full texts (Cohen's $\kappa = 0.76$), with disagreements resolved by discussion. The three search rounds are shown separately in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram (Figure S1).

Because the reported metrics, tasks, datasets, and model versions were highly heterogeneous, the synthesis was primarily descriptive. An extracted metric record was defined as one numerical estimate for one outcome reported under a distinct evaluation condition, determined by study, review stage, task or dataset, model and version, prompting condition, comparator, and, when separately reported, run. Thus, one study could contribute multiple records. Within each outcome family, records were summarized using the median, interquartile range (IQR), and full observed range, with numbers of contributing studies and records reported separately. Prompt strategy was coded for the corresponding evaluation condition and analyzed descriptively. Random-effects pooling was secondary and exploratory and was limited to estimates with variance information. Further details are reported in the Supplementary Materials (Tables S2–S6).

After deduplication and screening, 168 studies were included. In tasks requiring semantic comprehension (title-and-abstract and full-text screening), LLMs showed their most favorable, though still uneven, results. Across extracted performance-metric records, the median recall (sensitivity) was approximately 0.88 (IQR 0.67–0.97), but individual values spanned the full 0–1 range; specificity (median = 0.87, IQR 0.66–0.96) and precision (median = 0.75, IQR 0.34–0.92) were lower and more variable, reflecting frequent over-inclusion of irrelevant citations (Table S5). High recall is desirable because missing a relevant study is a critical error, and several studies reported recall comparable to that of human reviewers.[8,9] These findings support a role for LLMs as an additional screener that flags clearly irrelevant citations for human confirmation, rather than as an autonomous decision-maker. In exploratory prompt subgroups analysis, zero-shot prompting remained most common. One- or few-shot prompting combined with chain-of-thought had the highest median recall (0.97), whereas zero-shot chain-of-thought had a lower median recall (0.72); because the subgroups were small, heterogeneous, and

unadjusted, no consistent causal advantage should be attributed to a specific prompting strategy (Table S6).[10]

In data extraction, LLMs structured information from clinical-trial reports with a median accuracy of approximately 0.81 and a median F1 score of approximately 0.81. Performance was generally adequate for standardized text elements such as study design, country, and sample size, at times comparable to that of junior reviewers, but declined for numerical data drawn from complex tables or non-standardized text.[11,12] Reported errors included the mis-mapping of table rows and columns and confusion between standard deviations and standard errors when labeling was implicit. Human verification therefore remains indispensable for statistical data extraction, even where qualitative study characteristics can be reliably automated.

By contrast, LLMs were least reliable in literature searching and in quality assessment. In the search stage, models frequently produced hallucinations, fabricating non-existent citations or misapplying Boolean logic, and achieved recall well below that of strategies designed by information specialists.[13,14] Some teams translated research questions into search strings with moderate success, but the outputs were variable and difficult to reproduce. For quality assessment, agreement with human appraisers was modest and inconsistent: across extracted consistency-metric records, the median Cohen's κ was approximately 0.64 (IQR 0.43–0.83; range −0.16 to 1.00), and several trials reported κ below 0.4.[15,16] LLMs could identify explicit methodological statements, for example, that randomization was performed, but struggled with the inferential judgment needed to detect subtle or incompletely reported sources of bias.

Efficiency gains were substantial but heterogeneous. Reported reductions in time varied widely (median fold-reduction of approximately 4; IQR 0.73–33.69), and individual reports described screening tasks shrinking from days to hours.[13,17] Work saved over sampling and related workload metrics had a median of approximately 0.64 (IQR 0.40–0.80). LLMs generally showed higher internal consistency (reproducibility across repeated runs of the same prompt) than agreement with human reviewers, and this agreement varied with both the prompt strategy and the specific model version[18]. Newer model versions tended to perform better, but frequent updates complicate longitudinal comparison: a prompt that performs well on one version may behave differently on another, undermining the stability of review protocols.

Several issues warrant caution. Hallucination is a critical risk, particularly the fabrication of references or data points absent from the source text, which can corrupt a meta-analysis. Some models also tend to be overly agreeable, or to overly aggressive exclusion under prompts emphasizing workload reduction; both behaviors require careful prompt design to mitigate. The opacity of proprietary models further challenges reproducibility, because provider updates can alter the deployed model or its observed behavior over time. Transparent reporting is therefore essential: authors should record the exact model version, temperature and other settings, full verbatim prompts, and the extent of human verification. PRISMA 2020 remains the appropriate baseline for reporting systematic reviews. PRISMA-Transparent Reporting of Artificial Intelligence in Comprehensive Evidence Synthesis(PRISMA-trAIce), a recently published checklist designed specifically for reporting AI used as a methodological tool in systematic literature reviews, may provide useful additional items, although it is a community-derived proposal rather than a formal consensus extension.[19]

Taken together, current LLMs are best positioned as copilots rather than as autonomous replacements. A human-in-the-loop model is essential: LLMs can absorb the high-volume burden of screening and initial extraction, while human experts retain control of search-strategy design and final quality assessment. These findings broadly align with recent reviews, which also found the most promising performance in screening and more persistent limitations in searching and judgment-intensive appraisal.[3–5] Our incremental contribution is an updated quantitative synthesis that separates performance, consistency, efficiency, and prompt-strategy results. The wide performance ranges documented here mean that aggregate optimism can be misleading; performance should be verified within each individual review rather than assumed from summary statistics. This review has limitations: inclusion was restricted to English-language reports; the evidence base evolves faster than any single search can fully capture; the review was deliberately limited to five core stages and did not evaluate protocol registration, question formulation, evidence mapping, statistical synthesis, or report writing; and the heterogeneity of metrics, tasks, and model versions precluded firm pooled conclusions, so the quantitative summaries should be read as descriptive. Future work should pursue domain-specific fine-tuning, address unstable performance in multilingual reviews, and explore LLM-assisted living systematic reviews that screen new evidence as it appears. Practically, current evidence supports using LLMs primarily for review-specific, human-verified screening and first-pass extraction, whereas search-strategy design and final risk-of-bias judgments should remain human-led.

Table 1. Summary of Large Language Model Performance and Utility Across Systematic Review Stages

Review stage	Performance	Agreement/consistency evidence	Main challenges	Suggested role
Literature search	Low and highly variable; recall well below specialist-designed searches	Limited and generally poor; reproducibility concerns	Hallucinated citations; unreliable Boolean logic; poor reproducibility	Assistive only; human-led
Title and abstract screening	Moderate-to-high recall (median recall = 0.88)	Usually moderate, but highly variable	Over-inclusion; sensitivity to prompt design and topic complexity	Additional screener with human confirmation
Full-text screening	Moderate-to-high; sensitive to document complexity	Moderate in available studies; highly variable	Handling complex PDFs; nuanced exclusion criteria	Additional screener with human confirmation
Data extraction	Moderate-to-high for structured text (median accuracy = 0.81)	Moderate for structured fields; lower for complex numeric fields	Numeric and table errors; fabrication of missing values	Drafting extractor with human verification
Quality assessment	Low-to-moderate; inconsistent agreement (median κ = 0.64)	Low-to-moderate	Limited inferential reasoning; superficial appraisal	Assistive only; human-led

Note: Agreement descriptors summarize the direction and spread of reported agreement or replication metrics and are not a formal evidence grade; no prespecified thresholding scheme was applied. Values are descriptive medians across heterogeneous studies, tasks, and model versions; ranges are wide and between-study heterogeneity is high (see Supplementary Materials). Abbreviations: κ, Cohen's kappa; PDF, portable document format.

References

1. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*. 2017;7(2):e012545. doi:10.1136/bmjopen-2016-012545
2. Zhao WX, Zhou K, Li J, et al. A Survey of Large Language Models. *arXiv*. Preprint posted online November 24, 2023:arXiv:2303.18223. doi:10.48550/arXiv.2303.18223
3. 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 Inf Assoc*. 2025;32(6):1071-1086. doi:10.1093/jamia/ocaf063
4. Laignelot F, Martin GL, Ossman M, et al. Large language models show promising performance for some systematic review tasks but call for cautious implementation: a systematic review. *J Clin Epidemiol*. 2026;194:112221. doi:10.1016/j.jclinepi.2026.112221
5. Lieberum JL, Toews M, Metzendorf MI, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. *J Clin Epidemiol*. 2025;181:111746. doi:10.1016/j.jclinepi.2025.111746
6. Farhat F, Silva ES, Hassani H, et al. The scholarly footprint of ChatGPT: a bibliometric analysis of the early outbreak phase. *Front Artif Intell*. 2024;6:1270749. doi:10.3389/frai.2023.1270749
7. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. *arXiv*. Preprint posted online 2021. doi:10.48550/ARXIV.2108.07258
8. Oami T, Okada Y, Nakada T. Performance of a Large Language Model in Screening Citations. *JAMA Network Open*. 2024;7(7):e2420496. doi:10.1001/jamanetworkopen.2024.20496
9. Tran VT, Gartlehner G, Yaacoub S, et al. Sensitivity and Specificity of Using GPT-3.5 Turbo Models for Title and Abstract Screening in Systematic Reviews and Meta-analyses. *Annals of Internal Medicine*. Published online May 21, 2024. doi:10.7326/M23-3389
10. Cao C, Sang J, Arora R, et al. Prompting is all you need: LLMs for systematic review screening. *medRxiv*. Preprint posted online June 3, 2024:2024.06.01.24308323. doi:10.1101/2024.06.01.24308323
11. Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods*. 2024;15(4):576-589. doi:10.1002/jrsm.1710
12. Sun Z, Zhang R, Doi SA, et al. How good are large language models for automated data extraction from randomized trials? *medRxiv*. Preprint posted online February 21, 2024:2024.02.20.24303083. doi:10.1101/2024.02.20.24303083
13. Adam GP, DeYoung J, Paul A, et al. Literature search sandbox: a large language model that generates search queries for systematic reviews. *JAMIA Open*. 2024;7(3):ooae098. doi:10.1093/jamiaopen/ooae098
14. 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
15. Woelfle T, Hirt J, Janiaud P, Kappos L, Ioannidis JPA, Hemkens LG. Benchmarking Human-AI Collaboration for Common Evidence Appraisal Tools. *medRxiv*. Preprint posted online April 22, 2024:2024.04.21.24306137. doi:10.1101/2024.04.21.24306137
16. Šuster S, Baldwin T, Verspoor K. Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. *Research Synthesis Methods*. 2024;15(6):988-1000. doi:10.1002/jrsm.1749
17. Ruksakulpiwat S, Phianhasin L, Benjasirisan C, et al. Assessing the Efficacy of ChatGPT Versus Human Researchers in Identifying Relevant Studies on mHealth Interventions for Improving Medication Adherence in Patients With Ischemic Stroke When Conducting Systematic Reviews: Comparative Analysis. *JMIR mHealth and uHealth*. 2024;12(1):e51526. doi:10.2196/51526
18. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst Rev*. 2023;12(1):72. doi:10.1186/s13643-023-02243-z

19. Holst D, Moenck K, Koch J, Schmedemann O, Schüppstuhl T. Transparent Reporting of AI in Systematic Literature Reviews: Development of the PRISMA-trAIce Checklist. *JMIR AI*. 2025;4:e80247. doi:10.2196/80247