Assessing the Effectiveness of Large Language Models in Automating Systematic Literature Reviews: Findings from Recent Studies

MSR155

Value Analytics Labs

Sumeyye Samur[1], Bhakti Mody[1], Rachael Fleurence[1,2,*], Elif Bayraktar[1], Turgay Ayer[1,3], Jagpreet Chhatwal[1,4]

1 Value Analytics Labs, Boston, MA, USA; 2 National Institutes of Health, DC, USA; 3 Georgia Institute of Technology, Atlanta, GA, USA; 4 Massachusetts General Hospital Institute for Technology Assessment, Harvard Medical School, Boston, MA, USA; *Dr. Fleurence contributed to this article in her personal capacity

BACKGROUND

- Systematic literature reviews (SLRs) are foundational for evidence synthesis in medical research but are often time-intensive and laborious. On average, completing a traditional SLR can take between 12 and 18 months and typically involves a team of at least three researchers.[1]
- Large Language Models (LLMs) offer potential for automating SLR tasks such as screening, data extraction, and bias assessment.
- However, the feasibility and performance of these models in conducting different steps of SLRs remain insufficiently documented.

OBJECTIVE

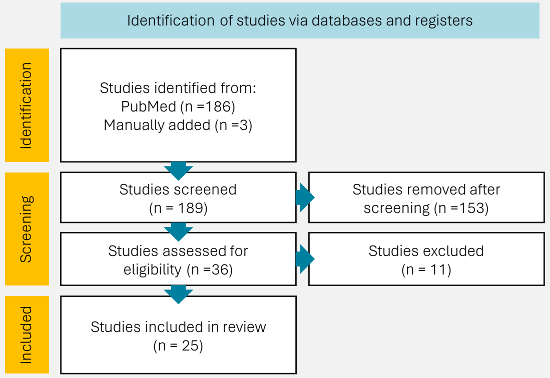
To evaluate the performance of LLMs across key tasks in the SLR process by reviewing recently published studies.

METHODS

We conducted a targeted literature review of studies applying LLMs in SLR workflows, focusing on performance metrics across multiple review tasks.

Study Identification: 25 studies published between Jan 2023 and Jan 2025 [2-26] (Figure 1)

Figure 1: *PRISMA Diagram for the Targeted Review of Studies applying LLM in SLR steps*



KEY FINDINGS

- LLMs show strong potential in automating key components of SLRs, particularly in early screening and data extraction.
- While different LLMs exhibit varying strengths and limitations across tasks, no study has comprehensively evaluated their performance across all SLR steps using AI.
- A hybrid approach that leverages the strengths of multiple LLMs may improve the overall efficiency and accuracy of SLRs.
- Performance of LLMs varies by task, suggesting that some steps in the SLR process may be more suitable for automation than others.
- Human oversight remains essential for conducting AI-assisted SLRs.
- Further research is needed to fully explore the capabilities and limitations of LLMs within SLR workflows.

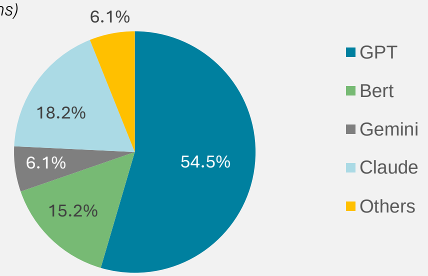
- **LLMs Assessed:** GPT, Gemini, Bert, Claude and others such as Google PaLM 2 and Meta Llama 2
- **SLR Tasks Analyzed:** Title and abstract screening, full-text screening, data extraction and bias assessment
- **Performance Metrics:** Sensitivity, specificity, accuracy, and inter-rater agreement (e.g., kappa score).

RESULTS

LLMs Assessed

Among all LLMs, GPT was the most frequently evaluated LLM across studies (Figure 2).

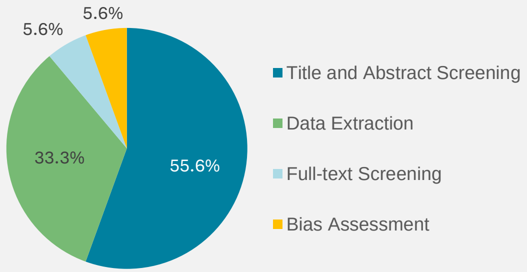
Figure 2: *Distribution of LLMs Evaluated (among 33 independent evaluations)*



Performance Evaluation

Among all SLR steps, title and abstract screening using LLMs was the most frequently evaluated one across studies (Figure 3).

Figure 3: *Distribution of SLR Tasks Evaluated (among 18 independent evaluations)*



LLMs demonstrated strong performance across various systematic review tasks (Table).

- **Title screening:** Sensitivity 94.3%–96.2%; specificity 85.5%–99.6%.
- **Abstract screening:** Accuracy 80%–97.5%; sensitivity 62%–95%; specificity 65%–98.7%.
- **Full-text screening:** Accuracy 87%; sensitivity 71.4%; specificity 93.8%.
- **Data extraction:** Accuracy 67%–96.3%; sensitivity 36%–96.2%; specificity >80%.
- **Bias assessment:** Strong agreement with human reviewers (kappa >0.89 for abstracts; 0.65 for full-text).

Table: *Performance of LLMs Across SLR Tasks*

Task	Best Performing Model	Metric	Performance Range
Title Screening	GPT-3.5	Sensitivity	94.3%–96.2% [18]
		Specificity	85.5%–99.6% [18]
Abstract Screening	GPT-4 (Acc, Spec) GPT-3.5 (Sens)	Accuracy	80%–97.5% [22, 26]
		Sensitivity	62%–95% [23, 26]
		Specificity	65%–98.7% [22, 23]
Full-Text Screening	GPT-4	Accuracy	87% [22]
		Sensitivity	71.4% [22]
		Specificity	93.8% [22]
Data Extraction	Claude 2 (Acc, Sens) GPT-4 (Spec)	Accuracy	67%–96.3% [14, 17]
		Sensitivity	36%–96.2% [14, 17]
		Specificity	>80% [17]
Bias Assessment	GPT-4	Kappa (Abstracts)	>0.89 [22]
		Kappa (Full-text)	0.65 [22]

REFERENCES

1. Systematic Reviews, https://belmont.libguides.com/systematicreviews, Accessed by April, 2025.
2. 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. Ann Intern Med 2024;177:791-799.
3. Akinseloyin O, Jiang X, Palade V. A question-answering framework for automated abstract screening using large language models. J Am Med Inform Assoc 2024;31:1939-1952.
4. Górska A, Tacconelli E. Towards Autonomous Living Meta-Analyses: A Framework for Automation of Systematic Review and Meta-Analyses. Stud Health Technol Inform 2024;316:378-382.
5. 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.
6. Gwon YN, Kim JH, Chung HS, et al. The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation. JMIR Med Inform 2024;12:e51187.
7. Khan MA, Ayub U, Naqvi SAA, et al. Collaborative Large Language Models for Automated Data Extraction in Living Systematic Reviews. medRxiv 2024.
8. Konet A, Thomas I, Gartlehner G, et al. Performance of two large language models for data extraction in evidence synthesis. Res Synth Methods 2024;15:818-824.
9. Oami T, Okada Y, Nakada TA. Performance of a Large Language Model in Screening Citations. JAMA Netw Open 2024;7:e2420496.
10. Treviño-Juarez AS. Assessing Risk of Bias Using ChatGPT-4 and Cochrane ROB2 Tool. Medical Science Educator 2024;34:691-694.
11. Woelfle T, Hirt J, Janiaud P, et al. Benchmarking Human-AI collaboration for common evidence appraisal tools. J Clin Epidemiol 2024;175:111533.
12. Ambalavanan AK, Devarakonda MV. Using the contextual language model BERT for multi-criteria classification of scientific articles. J Biomed Inform 2020;112:103578.
13. Teperikidis L, Boulmpou A, Papadopoulos C, et al. Using ChatGPT to perform a systematic review: a tutorial. Minerva Cardiol Angiol 2024;72:547-567.
14. Khraisha Q, Put S, Kappenberg J, et al. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods 2024;15:616-626.
15. Luo X, Chen F, Zhu D, et al. Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses. J Med Internet Res 2024;26:e56780.
16. Mutinda FW, Liew K, Yada S, et al. Automatic data extraction to support meta-analysis statistical analysis: a case study on breast cancer. BMC Med Inform Decis Mak 2022;22:158.
17. Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. Res Synth Methods 2024;15:576-589.
18. Matsui K, Utsumi T, Aoki Y, et al. Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews. J Med Internet Res 2024;26:e52758.
19. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev 2024;13:219.
20. Tao K, Osman ZA, Tzou PL, et al. GPT-4 performance on querying scientific publications: reproducibility, accuracy, and impact of an instruction sheet. BMC Med Res Methodol 2024;24:139.
21. Fabiano N, Gupta A, Bhambra N, et al. How to optimize the systematic review process using AI tools. JCPP Adv 2024;4:e12234.
22. Landschaft A, Antweiler D, Mackay S, et al. Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews. Int J Med Inform 2024;189:105531.
23. Issaiy M, Ghanaati H, Kolahi S, et al. Methodological insights into ChatGPT's screening performance in systematic reviews. BMC Med Res Methodol 2024;24:78.
24. Qin X, Liu J, Wang Y, et al. Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. J Clin Epidemiol 2021;133:121-129.
25. Aum S, Choe S. srBERT: automatic article classification model for systematic review using BERT. Syst Rev 2021;10:285.
26. Kohandel Gargari O, Mahmoudi MH, Hajisafarali M, et al. Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. BMJ Evid Based Med 2024;29:69-70

Contact: Sumeyye Samur, PhD, ssamur@valueanalyticslabs.com