# ASSESSING THE PERFORMANCE OF LARGE LANGUAGE MODELS IN AUTOMATING SYSTEMATIC LITERATURE REVIEWS: INSIGHTS FROM RECENT STUDIES

**MSR84**

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

[1] Value Analytics Labs, Boston, MA, USA, [2] National Institutes of Health, DC, USA, [3] Georgia Institute of Technology, Atlanta, Georgia, USA, [3] Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA,

## KEY FINDINGS

- LLMs demonstrate significant potential for automating key tasks in SLRs
- While they exhibit high accuracy and efficiency, their limitations, such as hallucinations and inconsistencies, underscore the need for human oversight to ensure the reliability of results.
- As LLM technology evolves, these models are likely to become indispensable tools, complementing human expertise and enhancing the efficiency of evidence synthesis.
- The current evidence underscores that these tools are best positioned as augmentative aids rather than replacements for human reviewers.

## BACKGROUND

- Large Language Models (LLMs) are increasingly being explored for use in systematic literature reviews (SLRs).
- Despite growing interest, their accuracy and reliability compared to human reviewers remain unclear.

## OBJECTIVE

Our objective was to summarize the findings from recent studies evaluating LLM performance in conventional SLR tasks.

## METHODS

We identified and reviewed studies assessing performance of LLM in a traditional SLR.

**Study Scope:** 13 studies (2023-2024) conducted across 8 countries: China, Japan, Hungary, Canada, Germany, Ireland, UK, and USA[1-13].

**Evaluated Tasks:** Abstract screening, Data extraction, Risk of bias assessment.

**LLMs Assessed:** ChatGPT, Claude 2, and others.

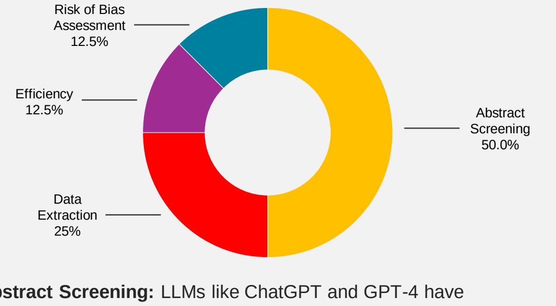
**Metrics Used:** Accuracy, Sensitivity, Specificity

## RESULTS

### Performance Evaluation

Recent studies evaluating LLMs demonstrated their potential to automate key SLR tasks, including abstract screening, data extraction, and risk of bias assessment, while improving efficiency and reducing costs. Most of the studies we reviewed focused on evaluating the performance of abstract screening with LLMs (Figure 1).

**Figure 1.** *Distribution of Studies Evaluating LLM Performance Across Various Criteria*




**Abstract Screening:** LLMs like ChatGPT and GPT-4 have demonstrated high accuracy in abstract screening, with some studies reporting accuracy rates exceeding 90%.[1,2]

## RESULTS (cont.)

**Data Extraction:** Models such as Claude 2 and GPT-4 have shown impressive data extraction capabilities, with accuracy rates often exceeding 96%.[3,4]

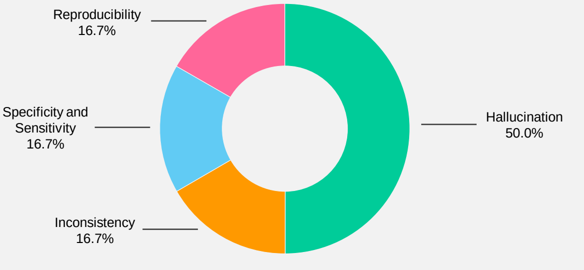
**Risk of Bias Assessment:** GPT-4 has been evaluated for its ability to assess the risk of bias, achieving a Cohen's kappa score of 0.90 when compared to human reviewers.[5-6]

**Efficiency:** The use of LLMs has been shown to significantly reduce the time required for tasks such as data extraction and abstract screening.[4,7]

### Limitations

Despite their promising performance, LLMs have notable limitations, with half of the studies highlighting concerns about hallucination generation (Figure 2)

**Figure 2.** *Distribution of Studies Highlighting Limitations of LLMs*




**Hallucination:** LLMs have a tendency to generate confident-sounding but fabricated responses, which can compromise the reliability of the results.[7-9]

**Inconsistency:** The performance of LLMs can be inconsistent, with different outputs for the same input, necessitating human oversight to validate the findings.[5]

**Specificity and Sensitivity:** While LLMs perform well in excluding irrelevant studies, their sensitivity in including relevant studies can be lower, potentially leading to the omission of important studies.[10]

**Reproducibility:** The reproducibility of results can be challenging due to the token limits and the need to segment texts, which may affect the coherence and accuracy of the extracted data.[4]

### Other Insights

**Complementary Role with Human Reviewers:** While some studies showcased Generative AI's ability to reduce human effort in SLRs, they emphasized the necessity of human involvement, particularly for final decision-making and verification. 4,5,7,10 Gartlehner (2023) found that combining human expertise with LLMs could enhance data extraction and synthesis accuracy.[3]

**Emerging Applications:** Beyond traditional SLR tasks, Luo (2024) explored the use of LLMs in defining research topics, generating statistical methods, and establishing inclusion/exclusion criteria, potentially broadening the utility of these models in SLRs and meta-analyses.[7]

## REFERENCES

[1] Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev. 2024 Aug 21;13(1):219. doi: 10.1186/s13643-024-02609-x. PMID: 39169386; PMCID: PMC11337893.

[2] Matsui K, Utsumi T, Aoki Y, Maruki T, Takeshima M, Takaesu Y. 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 Aug 16;26:e52758. doi: 10.2196/52758. PMID: 39151163; PMCID: PMC11364944.

[3] Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, Crotty K, Viswanathan M, Nussbaumer-Streit B, Booth G, Erskine N, Konet A, Chew R. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. Res Synth Methods. 2024 Jul;15(4):576-589. doi: 10.1002/jrsm.1710. Epub 2024 Mar 3. PMID: 38432227.

[4] Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. Pharmacoecon Open. 2024 Mar;8(2):205-220. doi: 10.1007/s41669-024-00476-9. Epub 2024 Feb 10. PMID: 38340277; PMCID: PMC10884375.

[5] Hasan B, Saadi S, Rajjoub NS, et al Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment BMJ Evidence-Based Medicine Published Online First: 21 February 2024. doi: 10.1136/bmjebm-2023-112597

[6] Landschaft, Assaf & Antweiler, Dario & Mackay, Sina & Rüping, Stefan & Wrobel, Stefan & Hoeres, Timm & Allende-Cid, Hector. (2024). Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews. International Journal of Medical Informatics. 189. 105531. 10.1016/j.ijmedinf.2024.105531.

[7] Luo X, Chen F, Zhu D, Wang L, Wang Z, Liu H, Lyu M, Wang Y, Wang Q, Chen Y. Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses. J Med Internet Res. 2024 Jun 25;26:e56780. doi: 10.2196/56780. PMID: 38819655; PMCID: PMC11234072.

[8] Jin Q, Leaman R, Lu Z. Retrieve, Summarize, and Verify: How Will ChatGPT Affect Information Seeking from the Medical Literature? J Am Soc Nephrol. 2023 Aug 1;34(8):1302-1304. doi: 10.1681/ASN.0000000000000166. Epub 2023 May 31. PMID: 37254254; PMCID: PMC10400098.

[9] Schopow N, Osterhoff G, Baur D Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review JMIR Med Inform 2023;11:e48933 URL: https://medinform.jmir.org/2023/1/e48933 DOI: 10.2196/48933

[10] Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. J Med Internet Res. 2024 Jan 12;26:e48996. doi: 10.2196/48996. PMID: 38214966; PMCID: PMC10818236.

[11] Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. 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 Jul;15(4):616-626. doi: 10.1002/jrsm.1715. Epub 2024 Mar 14. PMID: 38484744.

[12] Robinson, A., Thorne, W., Wu, B., Pandor, A., Essat, M., Stevenson, M., & Song, X. (2023). Bio-SIEVE: Exploring Instruction Tuning Large Language Models for Systematic Review Automation. ArXiv, abs/2308.06610.

[13] Kertész, Gábor and Czere, János Tibor and Zrubka, Zsombor and Gulácsi, László and Péntek, Márta, Towards Automating the Selection of Articles Reporting Eq-5d Data for Systematic Literature Reviews Using Large Language Models.

Contact: Jagpreet Chhatwal, PhD: jagchhatwal@mgh.harvard.edu