Use of Large Language Models (LLMs) for Full-text Screening in Systematic Literature Reviews: A Comparative Analysis

Rathi H[1,2], Malik A[1], Behera DC[1], Kamboj G[2]

[1]EasySLR, Gurugram, Haryana, India; [2]Skyward Analytics Pvt. Ltd., Gurugram, Haryana, India

MSR28

INTRODUCTION

- In healthcare, systematic literature reviews (SLRs) are increasingly being used to address research questions, leading to a rapid increase in the production of such reviews.[1] Based on our research, PubMed indexed 44,987 new SLRs in 2023 alone, and this number has increased more than five-fold over the last decade.
- As the need for SLRs continues to grow, the number of publications that need to be reviewed for review is also increasing at a similar rate. Despite advancements in technology over the past few decades, the production of SLRs has become more time-consuming and expensive than it was 35 years ago.[2-4]
- Large language models (LLMs) are advanced AI systems designed to understand and generate text that closely resembles human language.[5] LLMs can empower researchers to make more informed decisions and expedite SLRs.
- The objective of this study was to compare the performance of three LLMs — Anthropic Claude, OpenAI GPT, and our Proprietary model in the full-text screening stage of an SLR.

Key Takeaways

What is already known on this topic

- The potential of LLMs in automating various stages of the SLR process, such as literature search, screening, and data extraction has already been demonstrated.[6-8]
- Our previous study focused on comparing LLM performance in the primary screening (title-abstract) stage of an SLR.[9]

What this study adds

- This study adds to the existing knowledge by evaluating three LLMs (Anthropic Claude, OpenAI GPT, and our Proprietary model) in the full-text screening stage, comparing their performance against double screening by human reviewers.
- The findings highlight LLMs' potential to assist in the full-text screening stage of an SLR, with all three LLMs demonstrating similar performance in the analysis.

METHODS

- To conduct this comparative study, we implemented a standardized approach across all three LLMs, namely Anthropic Claude, OpenAI GPT, and our Proprietary model.
- We fed identical screening rules and search strategies to all three LLMs for full-text screening of 100 studies. These studies were previously identified and included in the primary screening level.
- The same 100 studies were screened by two independent human reviewers with a third reviewer reconciling any discrepancies.

Reference response

- The final decisions made by the human reviewers were taken as the reference response to assess the performance of the LLMs.

Evaluation metrics

- The evaluation metrics included decision match rate, precision, sensitivity score, specificity score, and F1 score (**Table 1**). These metrics provided a holistic assessment of the LLMs' performance in comparison to human reviewers.

Table 1. Evaluation metrics for performance analysis

Metric	Definition
Decision match rate	Cases where inclusion and exclusion decisions were identical between the human reviewer's final decision and LLM
Sensitivity (recall)	Proportion of included studies that are accurately predicted as included
Specificity	Proportion of excluded studies that are accurately predicted as excluded
Precision	Proportion of studies predicted as included that were actually included
F1 score	Harmonic mean of Precision and Sensitivity

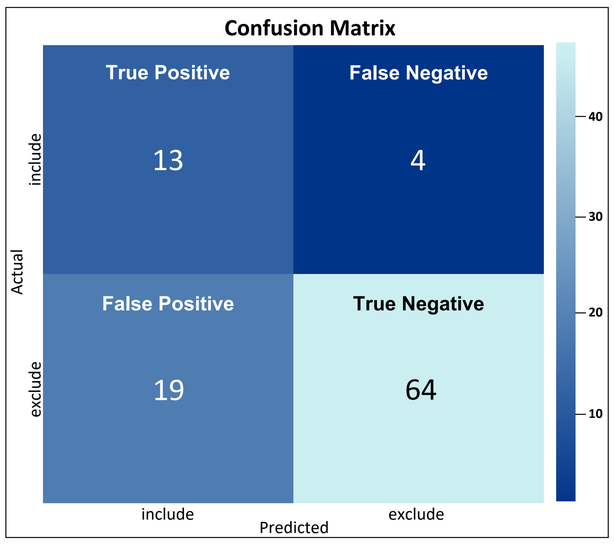
LLM: Large language model.

RESULTS

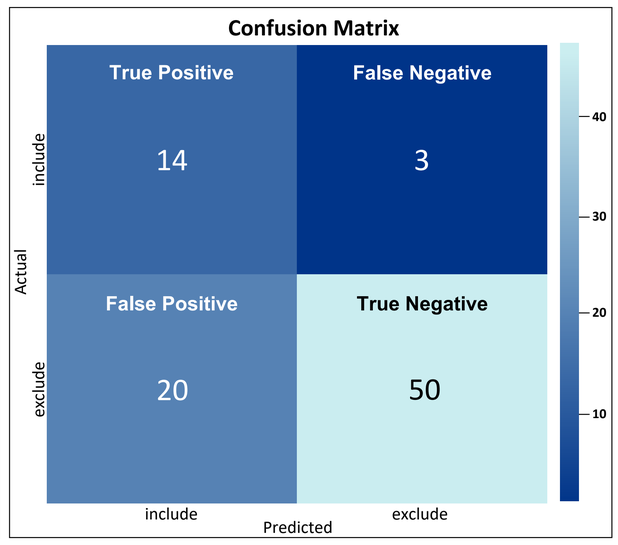
- The LLMs were evaluated using confusion matrices, which are tables that summarize the performance of a model by comparing actual and predicted classes (**Figure 1**).
- Anthropic Claude, OpenAI GPT, and our Proprietary Model scored a decision match rate of 77.0%, 73.6%,and 72.4%, respectively (**Table 2**).
- The corresponding sensitivity scores were 0.76, 0.82, and 0.94, with specificity scores being 0.77, 0.71, and 0.67, respectively.
- Anthropic Claude, OpenAI GPT, and our Proprietary Model had a similar precision of 0.40, 0.41, and 0.41, respectively.
- The corresponding F1 scores were 0.53, 0.55, and 0.57, respectively.

Figure 1. Confusion matrices for three LLMs*

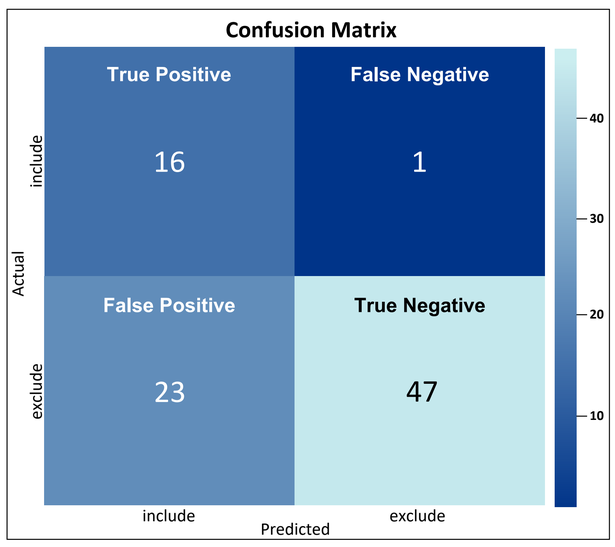
(a) Anthropic Claude



(b) OpenAI GPT



(c) Proprietary model



*In a confusion matrix:

1. 'True Positive' represents the number of instances where model correctly includes an actual included study
2. 'True Negative' represents the number of instances where model correctly excludes an actual excluded study
3. 'False Positive' is when the model wrongly includes an actual excluded study
4. 'False Negative' is when the model wrongly excludes an actual included study.

Table 2. Assessment outcomes for three LLMs

Metric	Anthropic Claude	OpenAI GPT*	Proprietary model*
Decision match rate	**77.0**	73.6	72.4
Precision	0.40	**0.41**	**0.41**
Sensitivity	0.76	0.82	**0.94**
Specificity	**0.77**	0.71	0.67
F1 score	0.53	0.55	**0.57**

**LLM was unable to take decisions for 13 studies.*
Text highlighted in 'blue' denotes the highest value of individual performance metric among the three LLMs.
LLM: Large language model.

Scenario analysis

- Further analysis revealed that the performance metrics of the LLMs varied significantly under different scenarios, particularly in response to changes in screening rules and the number of studies analyzed.
- This highlights the importance of considering various factors that may influence LLMs' performance in real-world applications.

CONCLUSIONS

- The findings highlight LLMs' potential to assist with the SLR process. All three LLMs were comparable in the decision match rate and F1 score metric. While in this simulation, our Proprietary Model showed a better sensitivity score and F1 score than Anthropic Claude and OpenAI GPT.
- These results should be interpreted cautiously, as they may vary with different research questions.
- Future research should consider analyzing the performance of LLMs on larger datasets and calibrating the framing of screening rules for better understanding by LLMs.
- Future analyses will delve into the utilization of LLMs in the process of data extraction.

FUNDING

No funding was received for this study.

REFERENCES

1. Norman C. Systematic review automation methods. Information Retrieval [cs.IR]. Université Paris-Saclay; Universiteit van Amsterdam. 2020; 2. Lau J. Systematic review automation thematic series. Systematic reviews. 2019;8:1-2; 3. Borah R et al. 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; 4. Tsafnat G et al. Systematic review automation technologies. Systematic reviews. 2014;3:1-5; 5. Tustumi F et al. Future of the language models in healthcare: The role of ChatGPT. Arq Bras Cir Dig. 2023;36:e1727; 6. Qureshi R et al. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation?. Systematic Reviews. 2023;12(1):72; 7. Guo E et al. Automated paper screening for clinical reviews using large language models. arXiv preprint arXiv:2305.00844. 2023; 8. Alshami A et al. Harnessing the power of ChatGPT for automating systematic review process: Methodology, case study, limitations, and future directions. Systems. 2023;11(7):351; 9. A Comparative Analysis of Large Language Models Utilised in Systematic Literature Review. ISPOR Presentations Database. Available at: https://www.ispor.org/docs/default-source/euro2023/eu-ispor-2023podium-presentationv1-0132550-pdf. Accessed on 5 April 2024.

Poster presented at ISPOR Annual International Meeting 2024, Atlanta, GA, USA (5-8 May 2024)