Transforming Systematic Literature Reviews: Unleashing the Potential of GPT-4, Cutting-Edge Large Language Model to Elevate Research Synthesis

MSR57

Sumeet Attri[1], Rajdeep Kaur[1], Barinder Singh[2], Pankaj Rai[1]
[1]Pharmacoevidence, Mohali, India, [2]Pharmacoevidence, London, UK

CONCLUSION

GPT-4 can potentially replace one of the reviewers in a standard systematic literature review, offering accuracy on par with subject matter experts. Further research is necessary to assess whether similar benefits extend to other language models and explore how prompt design variations could influence outcomes.

INTRODUCTION

- A systematic literature review (SLR) is a comprehensive and thorough survey of the most recent research literature on a topic or problem. However, the traditional method of carrying out systematic reviews is acknowledged to be laborious[1,2]
- The increasing integration of Artificial Intelligence (AI) in research is driven by its extensive potential, providing clear advantages over traditional methods
- It can handle large volumes of data efficiently, enhancing the overall speed and precision of the screening process
- Recent advances in AI, specifically the development of Large Language Models (LLMs) such as the generative pre-trained transformer (GPT) class models, have demonstrated the ability to generate and summarize text[3]

OBJECTIVE

- This study explores the efficiency of advanced language models, such as the generative pre-trained transformer (GPT-4), in automating the intricate procedures involved in systematic literature reviews

METHODS

- Embase®, Medline®, and Cochrane were utilized to identify relevant randomized controlled trials (RCTs) in patients with post-traumatic stress disorder
- The pre-defined inclusion/exclusion criteria were used to define the prompt used in the GPT-4, based on which GPT-4 decided whether to include or exclude each citation based on the title and abstract
- A subject matter expert (SME) with over ten years of experience in conducting SLRs optimized and refined the prompt based on the results obtained from a small subset of the citations, which was delivered through a Python API, to identify evidence aligning with crucial inclusion and exclusion criteria
- The optimized prompt was then applied to the complete dataset
- An evaluation was conducted to compare the screening results of AI and human reviewers, measuring their agreement levels and assessing the accuracy with which publications were identified for inclusion in the systematic literature review
- Figure 1 provides a comprehensive depiction of the entire process
- The performance of the model was assessed using True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) values from the confusion matrix, as shown in Figure 2, using the formula outlined in Table 1

RESULTS

- The study retrieved a total of 545 publications from biomedical databases. After removing duplicates, 519 publications were selected for screening based on their titles and abstracts.
- Automated screening using GPT-4 resulted in the inclusion of 17.73% of publications in comparison to 10.98% by a human reviewer
- The overall agreement, or the accuracy, between GPT-4 and the human reviewer stood at 90.17%, with reported sensitivity and specificity rates of 85.96% and 90.6%, respectively
- Both screening techniques identified all relevant publications however, the human reviewer required an additional five hours to complete the screening of the 519 publications

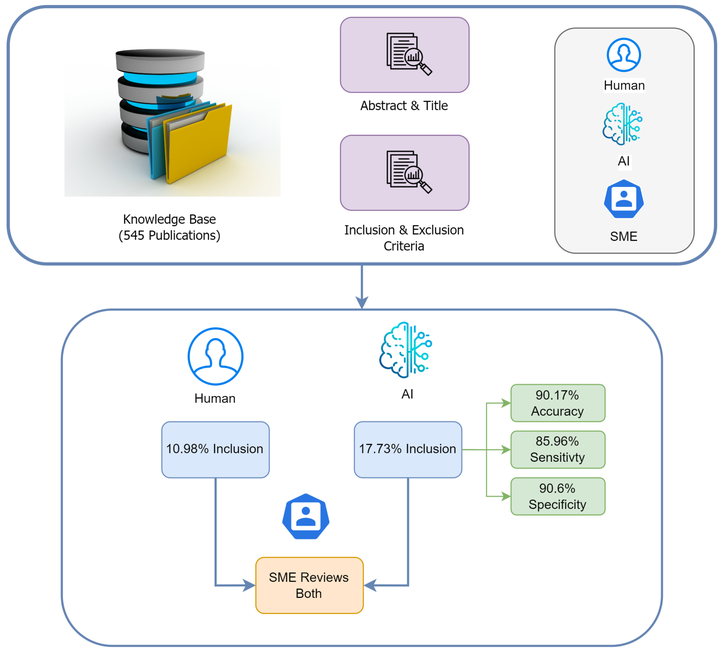

Figure 1: First stage screening using GPT-4

Table 1: Performance matrix

Accuracy =	$\frac{TP + TN}{TP + FP + TN + FN}$
Sensitivity =	$\frac{TP}{TP + FN}$
Specificity =	$\frac{TN}{TN + FP}$
Precision =	$\frac{TP}{TP + FP}$

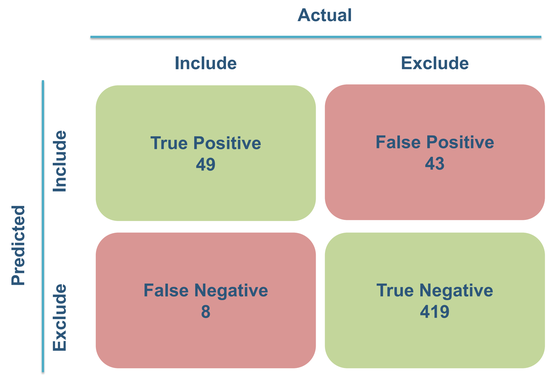

Figure 2: Confusion matrix

REFERENCES

1) Mahuli, S. A., Rai, A., Mahuli, A. V., & Kumar, A. (2023). Application ChatGPT in conducting systematic reviews and meta-analyses. *Br Dent J*, *235*(2), 90-92., 2) Issaiy, M., Ghanaati, H., Kolahi, S., Shakiba, M., Jalali, A. H., Zarei, D., ... & Firouznia, K. (2024). Methodological insights into ChatGPT's screening performance in systematic reviews. BMC Medical Research Methodology, 24(1), 78.; 3) Susnjak, T. Prisma-DFLLM: An extension of prisma for systematic literature reviews using domain-specific finetuned large language models. arXiv preprint arXiv: 230614905. 1–20. 2023. ArXiv, 1-20.

DISCLOSURE

SA, RK, BS, and PR, the authors, declare that they have no conflict of interest

For further queries, please contact: Barinder.Singh@Pharmacoevidence.com