Revolutionizing Systematic Literature Reviews: Harnessing the Power of Large Language Model (GPT-4) for Enhanced Research Synthesis

MSR15

Kaur R[1], Rai P[1], Attri S[1], Kaur G[1], Singh B[2]
[1]Pharmacoevidence, Mohali, India, [2]Pharmacoevidence, London, UK

Evidence Pharmaco

CONCLUSION

This investigation highlights the efficiency of GPT-4 over traditional SLR methods. Practically, attaining a 95% concurrence rate with a two-review human process is challenging. The outstanding accuracy of GPT-4, comparable to SME, suggests substituting one review of the traditional approach with a GPT-4 review to expedite the screening process. Future research should explore these benefits across language models and assess the impact of diverse prompts on outcomes

INTRODUCTION

- Systematic literature reviews (SLRs) are crucial for evidence-based decision-making and involve extensive literature searches and analysis. The procedure is time-consuming due to the enormous number of publications[1,2]
- The AI-based Large Language Model (LLM) can streamline the labor-intensive process of SLRs, leveraging its remarkable performance in its ability to quickly and accurately analyze large volumes of textual data using advanced natural language processing (NLP) algorithms
- The growing utilization of AI in the field of research is propelled by its widespread potential, offering distinct advantages over conventional methods of conducting SLRs
- AI notably reduces human errors and workload, increases productivity, ensures quick turnaround, and maintains consistency

OBJECTIVES

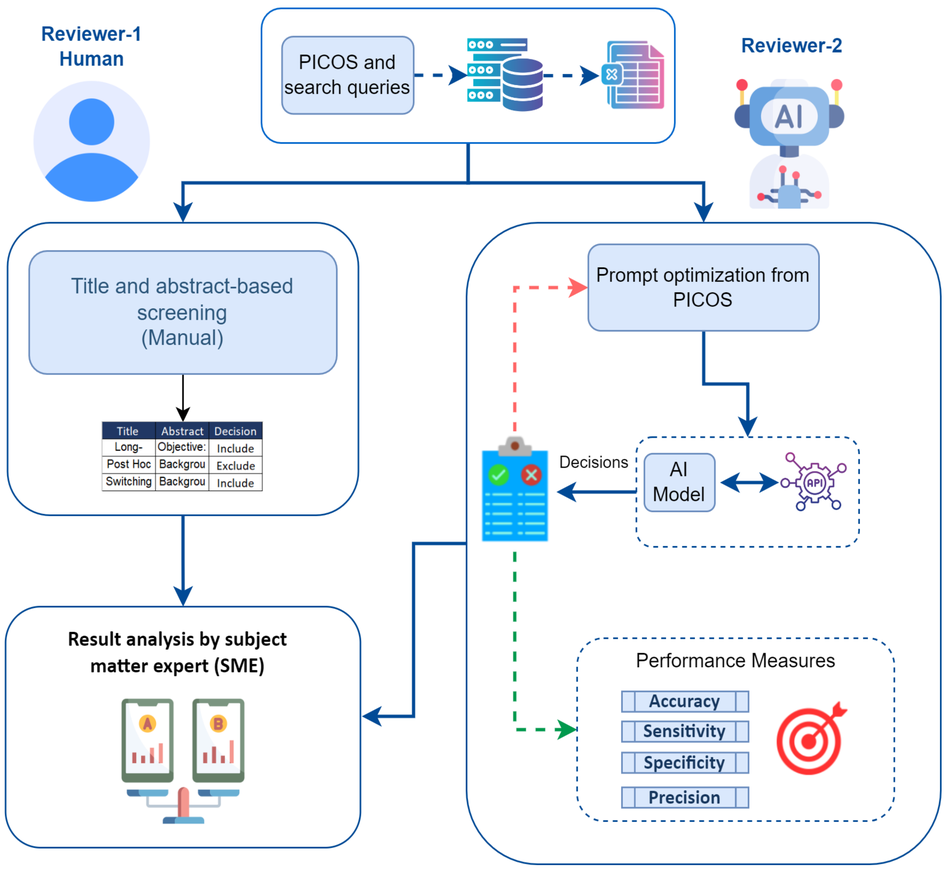
- This study specifically examines the capabilities of large language models, like a generative pre-trained transformer (GPT-4), in automating the complex processes of SLRs

METHODS

- Embase®, Medline®, and Cochrane were searched to identify relevant randomized controlled trials (RCTs) in patients with schizophrenia
- A subject matter expert (SME) with over a decade of experience in conducting SLRs optimized and fine-tuned the final prompt, delivered through a Python API to identify evidence meeting key inclusion and exclusion criteria
- Comparison of the screening results obtained via AI and human reviewer was conducted to evaluate agreement levels and assess the successful identification of publications incorporated in the final SLR

- A comprehensive description of the entire process is presented in Figure 1

Figure-1: First stage screening using GPT-4



- The model performance was evaluated using the confusion matrix which identifies the accuracy between actual and predicted values i.e.,
  - True Positive (TP)
  - False Positive (FP)
  - False Negative (FN)
  - True Negative (TN)

RESULTS

- Both a human reviewer and GPT-4 conducted the screening of 985 publications based on the titles and abstracts
- A total of 18.78% of publications were considered for inclusion by GPT-4 in comparison to 15.12% by a human reviewer
- Using predictive analytics, the overall agreement i.e., accuracy with GPT-4 and the human reviewer was 94.91%
- The sensitivity and specificity of GPT-4 were determined from the confusion matrix values depicted in Figure 2, yielding 95.30% and 94.85% respectively
- While both screening techniques identified all relevant publications, the human reviewer required an additional 10 hours to assess the 985 publications thoroughly

Figure 2: Confusion matrix to calculate performance matrices

	Actual Include	Actual Exclude
Predicted Include	TP: 142	FP: 43
Predicted Exclude	FN: 7	TN: 793

- Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$ = 94.91%
- Sensitivity = $\frac{TP}{TP+FN}$ = 95.30%
- Specificity = $\frac{TN}{TN+FP}$ = 94.85%
- Precision = $\frac{TP}{TP+FP}$ = 76.75%

REFERENCES

- 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.
- Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & Hadfield, K. (2023). Can large language models replace humans in the systematic review process? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *arXiv preprint arXiv:2310.17526*.
- Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E., & Zayed, T. (2023). Harnessing the power of ChatGPT for automating systematic review process: Methodology, case study, limitations, and future directions. *Systems*, *11*(7), 351.

For further queries, please contact: barinder.singh@pharmacoevidence.com

ISPOR Improving healthcare decisions