SA16

Automating systematic literature review abstract screening using Large Language AI Models: A development and validation study

Zdorovtsova N[1], Castanon A[1], Marsland A[1], Bray DB[1]

1 Health Analytics, Lane Clark & Peacock LLP. London, UK

Contact: Natalia.Zdorovtsova@lcp.uk.com

Summary

+ **Systematic literature reviews (SLRs) are known to be procedurally complex, time-consuming, and error-prone** despite their central role in evidence-based medicine[1].
+ The aim of the current study was to **compare the performance of two different LLMs on the accuracy of their abstract screening decisions** for three different medical SLR study questions. We used the PICOS (Population, Intervention, Comparison, Outcome, and Study Design) framework to construct LLM prompts that yield accurate inclusion and exclusion classifications in the context of SLR.
+ Our results point to the potential of **using LLMs to accelerate SLR timelines while retaining accuracy** at the abstract screening stage, particularly for SLRs of RCTs.

Background

- Abstract screening for SLR is costly and labour-intensive. In 2019, Michaelson and Reuter[2] estimated average yearly cost of all SLRs amounts to over $18 million for each academic institution, and over $16 million for each pharmaceutical company.
- Recently, Large Language Models (LLMs) have been explored as a promising means of automating the SLR process[3,4]. LLMs have demonstrated human-level performance on some SLR-related tasks, but questions remain about the comparative performance of different models, as well as how researchers should structure LLM prompts to guide accurate automated decision-making at the abstract screening stage.

Methods

- We developed LLM prompts for SLR abstract screening, drawing on the PICOS framework.
- The LLMs were prompted via API calls in Python to summarise each abstract, make inclusion/exclusion decisions, provide rationales, and report decision confidence levels (on a 1-5 scale).
- We validated the accuracy of this approach by comparing abstracts identified by the LLM for inclusion to those selected in previously-published SLRs which had been done manually[5,6,7]. Two of the SLRs focused on randomised-controlled trials (RCTs)[5,6], while the third focused on observational studies[7].
- For each SLR dataset, 20 runs were performed per LLM, and performance metrics were generated and compared across datasets and LLMs:

Classification accuracy: TP+TN/(TP+TN+FP+FN)
Positive Predictive Value: TP/(TP+FP)
Negative Predictive Value: TN/(TN+FN)
Sensitivity: TP/(TP+FN)
Specificity: TN/(TN+FP)

Step 1: SLR attrition data extraction

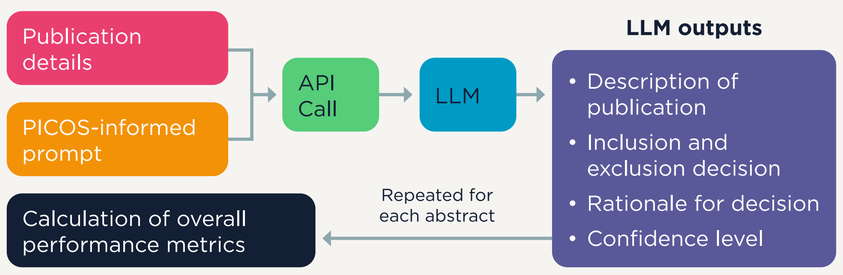
Used published PubMed search strings to identify abstracts for screening → Extracted human inclusion and exclusion criteria based on final SLR studies included → Compiled study titles, authors, journals of publication, dates of publication, full abstracts, and human inclusion/exclusion decisions

Step 2: PICOS-informed prompt construction

Each of our prompts included the following:

1. LLM instructions (information about the expected inputs and the desired output format [JSON])
2. The research question to be addressed in the SLR
3. A list of PICOS criteria for the SLR
4. A final instruction for the LLM to make inclusion and exclusion decisions very carefully based on the information provided

Step 3: LLM-based abstract screening



Results

- The two LLMs demonstrated high levels of accuracy (96.2%-96.9%) in classifying abstracts from SLRs of RCTs[5,6], and lower level of accuracy (37.0%-76.3%) for the SLR of observational studies[7].
- GPT-3.5 Turbo (see Figure 1) and GPT-4 Turbo (see Figure 2) achieved similar performance in terms of accuracy in classifying abstracts from SLR studies of RCTs, but GPT-3.5 Turbo was faster and lower-cost (see Table 1).
- Across all LLMs and SLR datasets, the false inclusion rate for classifications significantly exceeded the false exclusion rate.

Figure 1: GPT-3.5 Turbo Performance

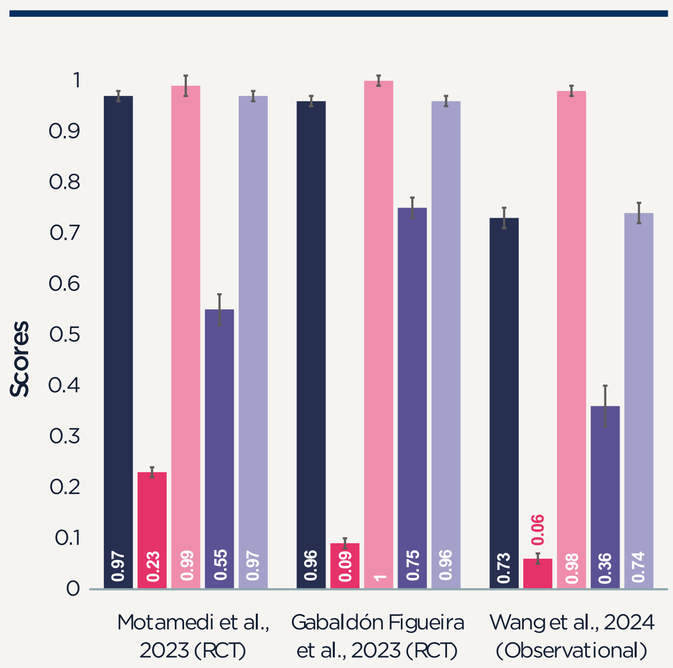


Figure 2: GPT-4 Turbo Performance

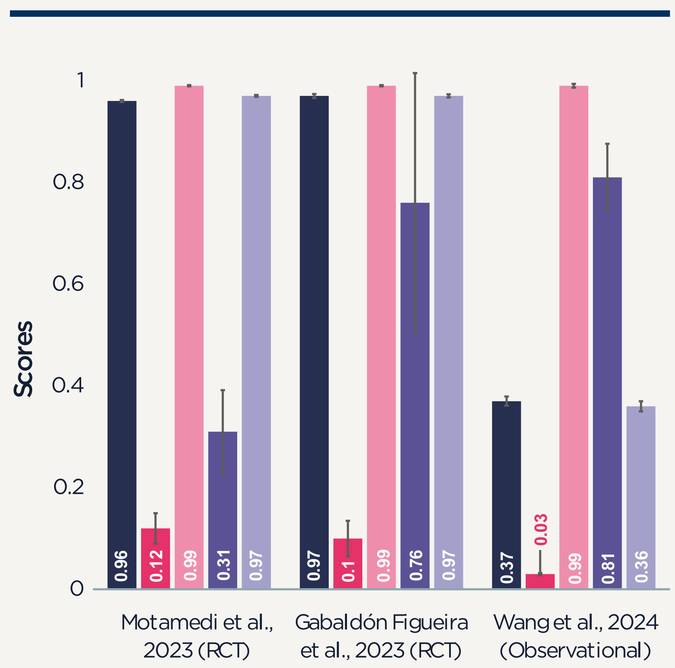


Table 1: Average costs and runtimes for GPT-3.5 Turbo and GPT-4 Turbo

		GPT-3.5 Turbo		GPT-4 Turbo	
SLR Dataset	No. Abstracts	Cost (USD)	Runtime (minutes)	Cost (USD)	Runtime (minutes)
Motamedi et al.[6] (RCT)	1,050	$1.16	8	$23.40	80
Gabaldón Figueira et al.[5] (RCT)	3,618	$3.97	24	$79.60	241
Wang et al.[7] (Observational)	540	$0.59	4	$11.88	40

Conclusions

- Overall, both GPT-3.5 Turbo and GPT-4 Turbo are able to perform abstract screening to a high degree of accuracy. However, **GPT-3.5 Turbo is significantly faster and cheaper**.
- One possible explanation for **lower classification accuracy in the case of observational SLR studies** is that for SLRs of RCTs, abstracts are typically excluded at the abstract screening stage, while in SLRs of observational studies, more abstracts are excluded later, during full-text screening. In the current study, we used a list of publication details derived from the abstract screening stage, but inclusion/exclusion labels from the full-text screening stage, which is further downstream in the SLR pipeline.
- Including LLM-based tools in the SLR workstream could **accelerate medical research consolidation**, but users must **review the reliability of abstract inclusion decisions and perform sensible technical checks**, as is already commonplace in traditional abstract screening procedures.

References:
1. Kolaski, K. et al. (2023). Guidance to best tools and practices for systematic reviews. DOI: 10.1186/s13643-023-02255-9
2. Michelson, M. et al. (2019). The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. DOI: 10.1016/j.conctc.2019.100443
3. Qureshi, R. et al. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation?. DOI: 10.1186/s13643-023-02243-z
4. Reason, T. et al. (2024). Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. DOI: 10.1007/s41669-024-00476-9
5. Gabaldón Figueira, J.C. et al. (2023). Topical repellents for malaria prevention. DOI: 10.1002/14651858.CD015422.pub2
6. Motamedi, M.A.K. et al. (2023). Local versus radical surgery for early rectal cancer with or without neoadjuvant or adjuvant therapy. DOI: 10.1002/14651858.CD002198.pub3
7. Wang, M. et al. (2024). Distribution of HPV types among women with HPV-related diseases and exploration of lineages and variants of HPV 52 and 58 among HPV-infected patients in China: A systematic literature review. DOI: 10.1080/21645515.2024.2343192