SWAR 66: Comparisons of the performance of large language models in review workflows versus human standards

Objective of this SWAR
To compare the performance of large language models (LLMs) in review workflows with human standards.

Study area: Title/abstract screening, Data extraction
Sample type: Review Authors, Consumers
Estimated funding level needed: Medium

Background
Automated large language models (LLMs) have shown promising capabilities for supporting systematic reviews. For example, the generalized pretrained transformer (GPT-)4 of openAI (https://openai.com/de-DE) can be effectively applied to title and abstract screening in clinical reviews, achieving performance levels close to human reviewers under certain conditions while offering substantial time savings.[1] Similarly, it has been reported that GPT-4 achieved 95% sensitivity and a negative predictive value of 99% in abstract screening, with a dramatic reduction in review time compared to human screening.[2]

Recent work has extended this potential to full-text screening. A workflow combining prompt engineering with retrieval-augmented generation (RAG) allowed full automation of both title/abstract and full-text screening, with GPT-4 achieving near-human accuracy: an article exclusion rate of 99.5%, specificity of 99.6%, sensitivity of 100%, and a 95.5% reduction in screening time.[3] At the same time, some limitations have been highlighted: GPT-4's accuracy is highly sensitive to dataset balance and prompt design, underscoring the need for robust validation before adoption.[4]

Following the release of GPT-5 in August 2025, initial studies indicate notable improvements in clinical reasoning, research tasks, and ethical evaluations compared to GPT-4. GPT-5 currently provides better performance compared to several other LLMs, while the access hurdle (in particular to the chat interface) is relatively low. However, responses generated via the chat-based user interface can vary considerably even when identical prompts are used. In light of such variability, the API offers several advantages, including improved reproducibility through version control, temperature settings, structured prompting workflows, and the automation of screening and extraction steps across large batches of studies. These features support standardized processes, minimize operator-driven variability, and increase transparency in model interactions. However, ultimately reproducibility and transparency lie with OpenAI. On the other hand, the families of the Large Language Models Meta AI (Llama) (https://www.llama.com/) offer clear advantages in terms of reproducibility, transparency, and governance as they are static LLMs. As open-source models whose version can be fully controlled and frozen, they support rigorous documentation and auditing, which is essential for scientific validity and compliance with evidentiary standards in systematic research. Moreover, local deployment allows full control over data handling, which is particularly relevant when working with copyright-protected or sensitive publication content.

This Study Within a Review (SWAR) [5] compares the use of both LLMs, i.e., GPT-5 (high performance) and Llama-70B (high reproducibility and transparency) for conducting methodological scoping reviews versus human standards. This comparative approach aims to assess the feasibility, accuracy, and efficiency of advanced language models as tools in review workflows.

Interventions and Comparators
Intervention 1: GPT-5 (high performance) for conducting methodological scoping reviews.
Intervention 2: Llama-70B (high reproducibility and transparency) for conducting methodological scoping reviews.
Intervention 3: Human standards (two reviewers) for conducting methodological scoping reviews.

Index Type:

Method for Allocating to Intervention or Comparator:

Outcome Measures
Primary: Sensitivity
Secondary: specificity, accuracy, interrater agreement, and screening efficiency

Analysis Plans
In order to evaluate the performance of the LLMs in comparison to human standards, sensitivity (recall), specificity, negative predictive value (NPV), positive predictive value (PPV/Precision), and accuracy will be calculated with their 95% confidence intervals (CIs). The interrater reliability (between the two reviewers; between the consensus agreement of the two reviewers and each of the LLMs) will be determined by Krippendorff's alpha with associated 95% CI. In addition, the unadjusted measure of overall percent agreement (pa) and the percent chance agreement (i.e., percent agreement expected by chance; pe) will be computed. The time required for title and abstract screening and data extraction by human reviewers will be compared with the time required by the LLMs to assess screening efficiency.

Possible Problems in Implementing This SWAR

References
1. Guo E, Gupta M, Deng J, et al. Automated paper screening for clinical reviews using large language models: data analysis study. Journal of Medical Internet Research 2024;26:e48996.
2. Issaiy M, Ghanaati H, Kolahi S, et al. Methodological insights into ChatGPT's screening performance in systematic reviews. BMC Medical Research Methodology. 2024;24(1):78.
3. Trad F, Yammine R, Charafeddine J, et al. Streamlining Systematic Reviews: A Novel Application of Large Language Models. arXiv preprint arXiv:241215247. 2024.
4. Khraisha Q, Put S, Kappenberg J, et al. 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:231017526. 2023.
5. Devane D, Burke NN, Treweek S, et al. Study within a review (SWAR). Journal of Evidence-Based Medicine 2022;15(4):328-32.

Publications or presentations of this SWAR design

Examples of the implementation of this SWAR

People to show as the source of this idea: Stella Erdmann
Contact email address: erdmann@imbi.uni-heidelberg.de
Date of idea: 01/08/2025
Revisions made by:
Date of revisions: