Optimising performance of generative artificial intelligence in systematic literature review screening using PICOS criteria

No. 10

when you see this icon, scan the code below for additional interactive content

Key findings: Automating screening using generative AI (GenAI) could speed up literature reviews, but performance may depend on model type, topic complexity, and prompt design.


Bernard Kerr
Prime, London, UK


George Miles
Prime, Knutsford, UK


Lucia Giles
Prime, Oxford, UK


Brian Norman
Prime, London, UK

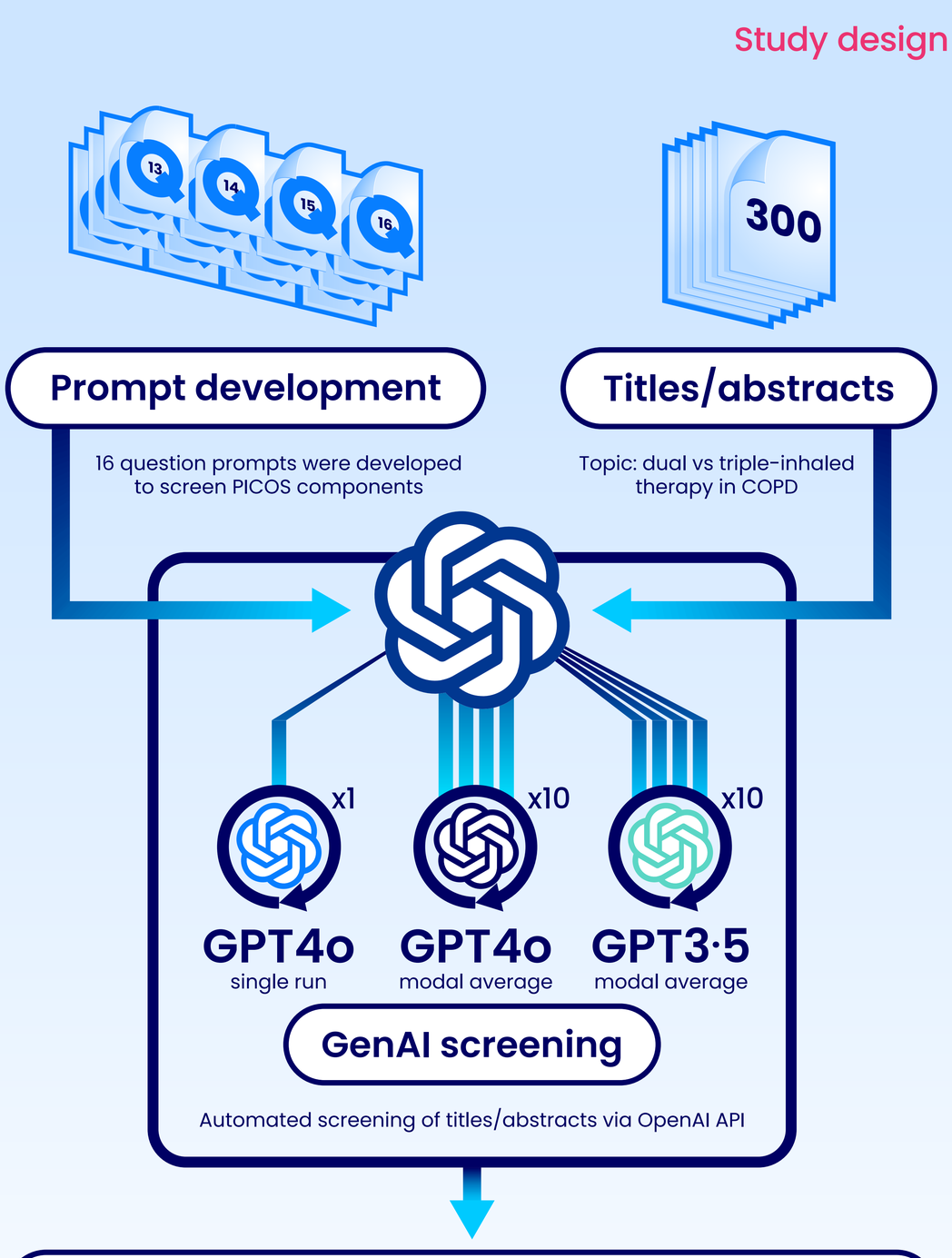

Gillian Sibbring
Prime, Knutsford, UK

Background

- Literature screening for SLRs is time-consuming.
- GenAI may streamline this process, but there are limited data on its performance.
- This study evaluated the level of agreement between genAI and a human reviewer when screening 300 titles and abstracts from a previous SLR (Figure 1).

Method

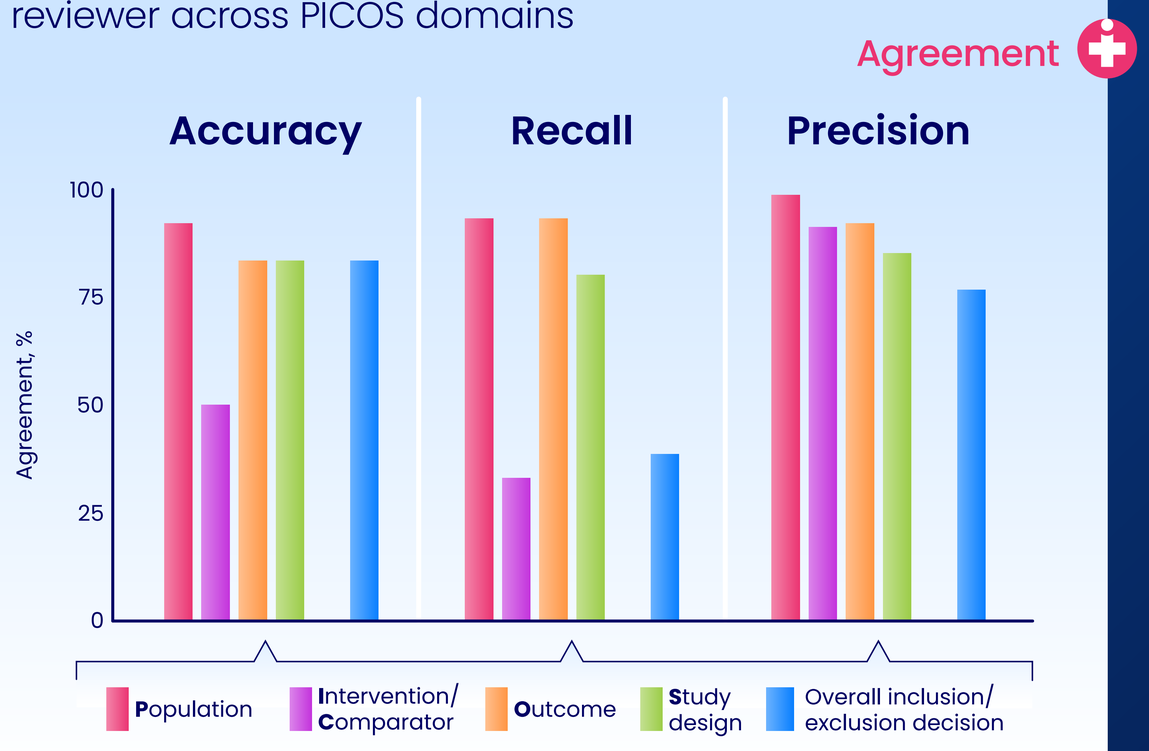
Figure 1. Approaches to assess the performance of genAI vs human screening of literature



Results

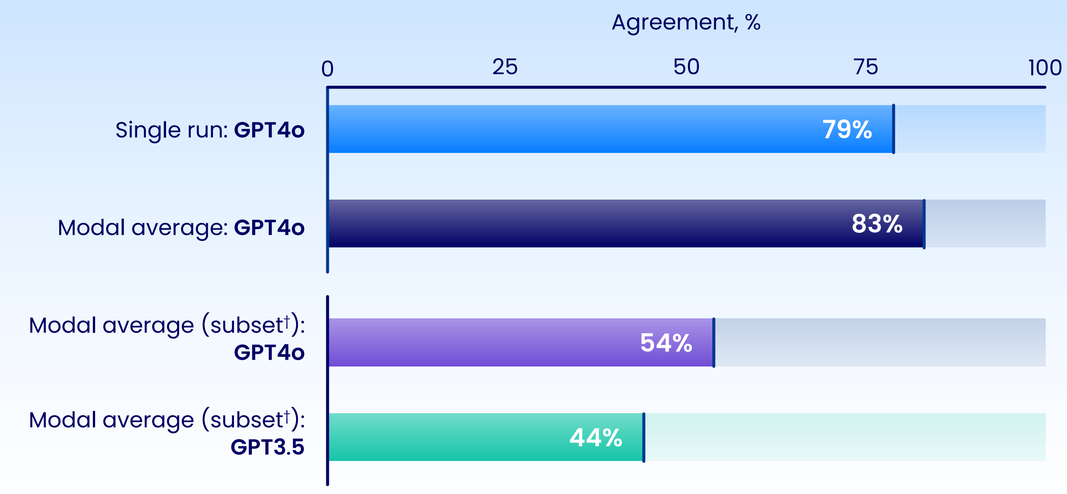
- GenAI showed high agreement with the human reviewer on Population, Outcome, and Study design, but low agreement on the Intervention/Comparator domain (Figure 2).
- Accuracy and precision were relatively high for the overall inclusion/exclusion decision, but recall was low.

Figure 2. Agreement between single-run GPT4o and human reviewer across PICOS domains



- Compared with the single-run approach, taking the modal answer of 10 runs with GPT4o resulted in slightly higher accuracy for the overall inclusion/exclusion decision (Figure 3).
- Conversely, accuracy was lower for the GPT3.5 subset than with the GPT4o subset.

Figure 3. Overview of accuracy for overall inclusion/exclusion decision across model scenarios



†In a selected subset of 50/300 records.

- Overall, agreement between genAI and the human reviewer (accuracy) was similar to what would be expected between two humans.[1]

Reference

1. Hanegraaf P, et al. *BMJ Open*. 2024;14(3):e076912.

Abbreviations: **AI**, artificial intelligence; **API**, application programming interface; **COPD**, chronic obstructive pulmonary disease; **GenAI**, generative AI; **PICOS**, population, intervention, comparator, outcome, study design; **SLR**, systematic literature review; **GPT**, generative pre-trained transformer.

Acknowledgements: The authors thank Lloyd Gwishiri, MD, for writing assistance; Adam Moriarty and Joey Cusano, for creative input; Richard McDonald for editorial support, as well as the Prime Digital and Production teams for their support with development of this poster. All opinions are the authors' own and do not necessarily reflect those of our employer.

Conflicts of interest: All authors are employees of Prime.

Presented at the European Meeting of ISMPP, London, January 2025.

18027519_Prime ISMPP GenAI Poster v5.indd 1

20/01/2025 09:14