Research Letter

AI-Assisted Systematic Review: Humans Still Need to Review All Abstracts for Inclusion

Hyelin Sung[1*], BSCN; Deyana Altahsh[1*], BSc; Scott Garrison[2], BASc, MD, PhD

[1]Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB, Canada
[2]Department of Family Medicine, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB, Canada
*these authors contributed equally

Corresponding Author:

Scott Garrison, BASc, MD, PhD
Department of Family Medicine, Faculty of Medicine & Dentistry
University of Alberta
6-60 University Terrace
Edmonton, AB T6G 2T4
Canada
Phone: 1 (587) 785 3012
Email: scott.garrison@ualberta.ca

Abstract

Although a general purpose (GPT-5), and a fine-tuned (ASReviewLab) artificial intelligence were able to rank abstracts for likely inclusion in a variety of Cochrane systematic reviews, some actually included studies were not highly ranked, necessitating human review of all abstracts.

JMIR Form Res 2026;10:e82896; doi: 10.2196/82896

Keywords: title and abstract screening; artificial intelligence; AI; large language model; LLM; machine learning; ChatGPT; GPT-5; ASReviewLab; systematic review

Introduction

Although systematic reviews (SRs) are the gold standard of evidence synthesis, they are time consuming, requiring evaluation of thousands of abstracts [1]. Artificial intelligence (AI) has the potential to identify and rank relevant studies for inclusion [2,3] but do they reliably rank all eligible abstracts high enough that only a subset of high-ranked abstracts require human review? We evaluated where studies actually included in published SRs fell within the "most likely for inclusion" rankings of a general-purpose AI (GPT-5), and an AI fine-tuned to identify SR-eligible studies (ASReviewLab)—assessing AI ranking reliability under worst-case constraints.

Methods

Overview

The most recently published 25 Cochrane SRs with ≥5 included studies were identified (as of May 5, 2025, from all subject areas), and their reported search strategies were replicated using the same databases, algorithms, and date ranges reported in each SR. We did not carry out gray literature searches, nor search databases with undefined algorithms, and limited abstracts to English. Deduplication and data export were completed using Covidence for each SR individually. Studies included in the published reviews were separated into "main results" publications and "supplementary studies" (eg, protocols, interim findings, subgroup analyses). Each publication record was separately ranked. We preidentified the AI tool as having "high-utility" if all eligible studies were placed/presented in the top 500, or top 15%, of AI-ranked abstracts (anticipated to shorten human review time by 85%).

For GPT-5 Thinking (version August 7, 2025), an AI prompt was iteratively developed to rank abstracts in order of likely inclusion. The prompt included 1) assessing whether the abstract reported on a randomized trial, 2) each SR's PICO (Patient/Population, Intervention, Comparison, and Outcome) criteria (setting, population, intervention, comparator, outcomes), and 3) an example of an included study (ie, the main results abstract for the trial with largest enrollment). An example of the final (best performing) prompt, used for all rankings, is provided in Multimedia Appendix 1.

For ASReviewLab (version 2.1.1), due to platform differences, PICO criteria could not be provided. Rather, per this AI's normal procedure of continually learning from human feedback, we provided it three included abstracts (those with highest enrollment), and subsequently labeled studies as included, or excluded, in the order ASReviewLab presented them. Given this interaction was time-consuming, we limited evaluation of ASReviewLab to the 10 most recent Cochrane SRs, and evaluated a maximum 500 studies for inclusion per SR.

Ethical Considerations

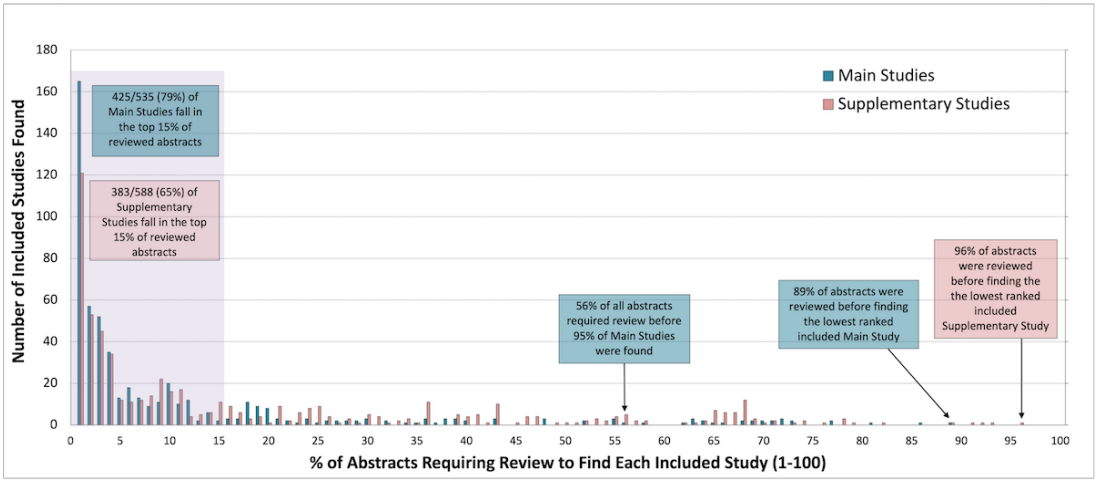
All studies were identified from previously published and publicly available SRs. Therefore, no research ethics approval was required. Ethical AI use was upheld through reporting of prompts and processes for replicability, and all AI outputs were evaluated by human reviewers for accuracy.

Results

Five SRs were excluded for having <5 included studies (ie, 25 of the 30 most recent SRs were eligible). The total number of abstracts requiring review ranged from 152 to 39,132 studies.

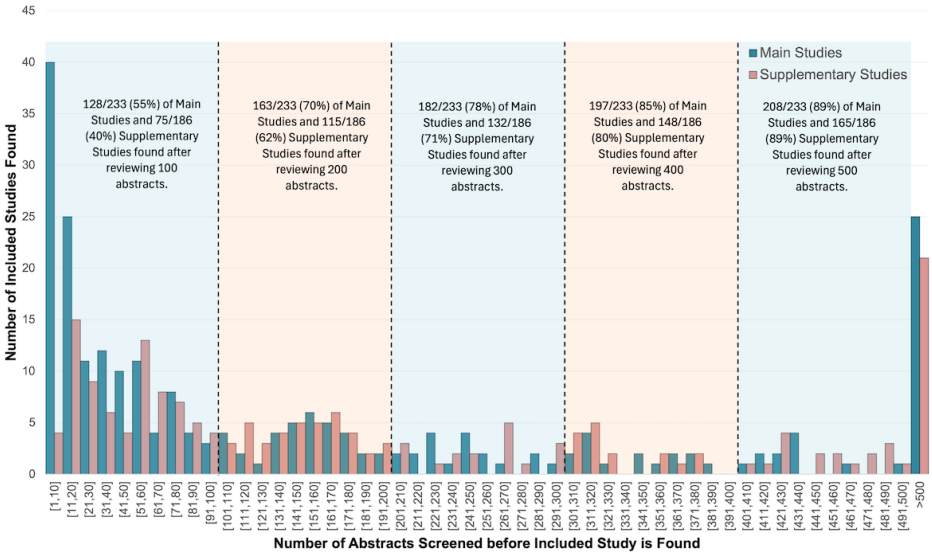
For GPT-5 Thinking, 144,120 abstracts were screened, of which 1123 were included studies—comprising 535 main results publications and 588 supplementary studies. Although 79.4% (n=425/535) of main results and 65.1% (n=383/588) of supplementary studies were within the highest-ranked 15% of abstracts, 89% (n=128,266/144,120) of abstracts were more highly ranked than the lowest-ranked main results publication, and 96% (n=138,355/144,120) of abstracts were more highly ranked than the lowest-ranked supplemental study (Figure 1).

Figure 1. GPT-5: % of abstracts requiring review to find each included study.



For ASReviewLab, 80,298 abstracts required review, of which 419 were included studies—comprising 233 main results publications and 186 supplementary studies. Although 89% of both main results (n=208/233) and supplementary studies (n=165/186) were identified in the first 500 AI-ranked abstracts, that left 11% (n=46/419) of included studies yet to be identified (Figure 2).

Figure 2. ASReviewLab: number of abstracts screened before included study is found.



Discussion

Both AI tools ranked the majority of included studies within the first 500, or first 15% of abstracts. However, many included studies required substantially more abstracts to be reviewed before they were identified. For GPT-5, where it was possible to see where all abstracts were ranked, some included studies would have required 96% of abstracts to be reviewed before all included studies were found – making these tools unreliable, given ALL eligible studies need to be identified (ie, unacceptable tail risk).

Our findings are limited to the two AI tools utilized and to SRs with well formulated PICO questions, as is the norm for the Cochrane Collaboration. It is possible that better prompts could have been developed to improve GPT-5 performance. Had we reviewed more than 500 abstracts with ASReview-Lab, we might also have found all included studies to lie within a still reasonable number of abstracts for humans to review in a timely manner – especially since reviewing more abstracts could have improved the AIs ability to recognize eligible studies. For GPT-5, there appeared to be no problem with prompt limits, domain variation, or publication types – as both main studies and supplementary studies had similar proportions of low-ranked included studies. Why some included studies received a low ranking is unclear.

Our findings align with other work suggesting AI models have potential to automate the SR process. However, the AI tools reported on here are not yet reliable enough to limit the number of abstracts human reviewers must evaluate if the goal is to find all eligible studies [4].

Our study is novel in providing a PICO-criteria-based prompt to GPT 5-Thinking, and in replicating abstract searches for multiple Cochrane reviews in order to identify all studies which should have been highly-ranked by AI tools. These Cochrane reviews additionally spanned many areas of medicine – from neonatal acupuncture to improving professional practice.

Unless performing a scoping review, where some relevant studies might be expected to be missed [5], the AI tools we examined are not yet sufficiently reliable at identifying eligible studies that lower-ranking studies do not require evaluation by a human reviewer. Doubtless this will change in time, as AI models evolve, for which further such evaluations of the ability to rank eligible studies would be worthwhile—including supplementary studies, and SRs of a non-medical nature, which pose different classification challenges.

Funding

Funding came exclusively in the form of Research Summer Studentship Awards to Hyelin Sung and Deyana Altahsh from the University of Alberta Deptartment of Family Medicine.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Sample prompt to rank abstracts in order of most to least likely for inclusion.
[DOCX File (Microsoft Word File), 26 KB-Multimedia Appendix 1]

References

1. Nussbaumer-Streit B, Ellen M, Klerings I, et al. Resource use during systematic review production varies widely: a scoping review. J Clin Epidemiol. Nov 2021;139:287-296. [doi: 10.1016/j.jclinepi.2021.05.019] [Medline: 34091021]
2. Affengruber L, van der Maten MM, Spiero I, et al. An exploration of available methods and tools to improve the efficiency of systematic review production: a scoping review. BMC Med Res Methodol. Sep 18, 2024;24(1):210. [doi: 10.1186/s12874-024-02320-4] [Medline: 39294580]
3. Fabiano N, Gupta A, Bhambra N, et al. How to optimize the systematic review process using AI tools. JCPP Adv. Jun 2024;4(2):e12234. [doi: 10.1002/jcv2.12234] [Medline: 38827982]
4. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing US closer to systematic review automation? Syst Rev. Apr 29, 2023;12(1):72. [doi: 10.1186/s13643-023-02243-z] [Medline: 37120563]
5. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, McEwen SA. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. Res Synth Methods. Dec 2014;5(4):371-385. [doi: 10.1002/jrsm.1123] [Medline: 26052958]

Abbreviations

AI: artificial intelligence
PICO: Patient/Population, Intervention, Comparison, and Outcome
SR: systematic review

Edited by Javad Sarvestan; peer-reviewed by Gaurav Kumar Gupta, Hassaporn Thongdaeng, Rohit Mamidipally; submitted 23.Aug.2025; accepted 18.Feb.2026; published 19.Mar.2026

Please cite as:
Sung H, Altahsh D, Garrison S
AI-Assisted Systematic Review: Humans Still Need to Review All Abstracts for Inclusion
JMIR Form Res 2026;10:e82896
URL: https://formative.jmir.org/2026/1/e82896
doi: 10.2196/82896

© Hyelin Sung, Deyana Altahsh, Scott Garrison. Originally published in JMIR Formative Research (https://formative.jmir.org), 19.Mar.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.