Performance of AI Tools for Automated Literature Screening in Evidence Synthesis

Minghao Ruan, Junhao Fan

Abstract

Background Literature screening constitutes a critical component in evidence synthesis; however, it typically requires substantial time and human resources. Artificial intelligence (AI) has shown promise in this field, yet the accuracy and effectiveness of AI tools for literature screening remain uncertain. This study aims to evaluate the performance of several existing AI-powered automated tools for literature screening.

Methods This diagnostic accuracy study employed a cohort to evaluate the performance of five AI tools—ChatGPT 4.0, Claude 3.5, Gemini 1.5, DeepSeek-V3, and RobotSearch—in literature screening. We selected a random sample of 1,000 publications from a well-established literature cohort, with 500 randomized controlled trials (RCTs) and 500 non-RCTs. Diagnostic accuracy was measured using the false negative fraction (FNF), false positive fraction (FPF), screening time, and the redundancy number needed to screen.

Results We reported the FNF for the RCT group and the FPF for the non-RCT group. RobotSearch exhibited the lowest FNF (6.4%; 95% CI: 4.6%–8.9%), whereas Gemini had the highest (13.0%; 95% CI: 10.3%–16.3%). The four large language models achieved FPF values between 2.8% and 3.8%, substantially lower than RobotSearch (22.2%). Mean screening times per article were 1.3 s (ChatGPT), 6.0 s (Claude), 1.2 s (Gemini), and 2.6 s (DeepSeek).

Conclusions The AI tools demonstrated commendable performance in literature screening but are not yet suitable as standalone solutions. They are best used as effective auxiliary aids, with hybrid human–AI screening offering the potential to improve both efficiency and accuracy.

1 Introduction

Evidence synthesis, often referred to as research synthesis, is an essential methodology for aggregating evidence from existing studies to inform healthcare decision-making [1]. Due to its inherent characteristics, the primary objective of evidence synthesis is to address disagreements and delineate what is known and unknown regarding a specific topic [2]. As its value has gained increasing recognition, evidence synthesis has emerged as a fundamental component of the contemporary medical paradigm, producing evidence through these syntheses and supporting the development of clinical guidelines and healthcare policies [3].

In evidence synthesis practice, standardized procedures are crucial for safeguarding against potential biases (e.g., publication bias) and errors (e.g., human error), thereby preventing adverse effects on the validity of the summarized evidence [4]. A well-conducted evidence synthesis requires meticulous attention to detail at each step, making the process often time- and labor-intensive [5–7]. Generally, high-quality systematic reviews typically take between 0.5 and 2 years to complete [8], raising concerns about efficiency in real-world scenarios where rapid and decisive healthcare decisions are often needed, especially during emergencies.

Literature screening is the most time-consuming process in evidence synthesis practice, particularly in the medical field [9–11]. The process generally involves two standardized steps: the initial screening of titles and abstracts, followed by full-text review. Each step requires two researchers or research groups working independently on the same task (i.e., double screening) [12, 13]. The first step aims to eliminate records that are clearly outside the scope of the research question, thereby reducing the workload associated with full-text screening [14]. When researchers specify a particular study design (e.g., randomized controlled trials, RCTs) relevant to a topic, screening titles and abstracts becomes more manageable because this information is usually explicitly reported [10]. This characteristic facilitates the development of artificial intelligence (AI) systems for literature screening, with the potential to reduce human workload and improve efficiency [15–17].

Currently, AI-powered literature screening tools can be broadly categorized into two types: semi-automatic tools (e.g., Rayyan) and fully automatic tools (e.g., RobotSearch) [10, 16–18]. Semi-automatic tools estimate the probability that each publication should be included or excluded, while the final decision remains with human reviewers. In contrast, fully automatic tools complete the screening process independently using algorithmic decision-making without human involvement [19–21]. Although fully automatic tools are expected to substantially improve the efficiency of evidence synthesis, their reliability remains an important concern. Specifically, it is unclear whether their screening decisions are sufficiently accurate for practical use. To address this question, we compared the performance of several currently available fully automatic AI-powered literature screening tools. Our findings provide empirical evidence regarding the feasibility of applying AI-powered tools to automated literature screening.

2 Methods

2.1 Study design

This study is a diagnostic accuracy study with a cohort, where the population refers to the body of literature being screened. Briefly, a cohort of literature was established and reviewed by human reviewers following standardized procedures. The literature was then categorized into two groups: those that were randomized controlled trials (RCTs) and those that were not, with equal group sizes obtained using simple random sampling. Using these RCT and non-RCT samples, the diagnostic accuracy of AI-powered tools was evaluated and compared.

This study focused specifically on the classification of RCTs versus other study types. An RCT was defined as a study in which participants were explicitly randomized into two or more groups, each receiving different interventions, as determined from the full text of the study [22]. The study was conducted in accordance with the STARD (Standards for Reporting Diagnostic Accuracy Studies) guidelines, where applicable [23].

2.2 Literature cohort

The literature cohort for this study comprised 8,394 retractions sourced from the Retraction Watch database up to April 26, 2023. Two experienced clinical epidemiology methodologists independently screened the exported records following standardized procedures.

The initial screening involved a review of titles and abstracts, with only those records excluded by both reviewers being discarded. In the second stage, the remaining records—categorized as "definite," "maybe," or "controversial"—were assessed using full texts. Any discrepancies encountered during full-text screening were resolved through discussion with a third senior methodologist. The Rayyan application (`https://www.rayyan.ai/`) was used for literature screening without employing its AI ranking system.

Following literature screening, 779 retractions were identified as RCTs, while 7,595 were classified as non-RCTs. To balance the sample sizes, a random sample of 500 articles was drawn from each group.

2.3 AI-powered tools for literature screening

The current study evaluated the performance of five AI-powered tools: RobotSearch, ChatGPT 4.0, Claude 3.5, Gemini 1.5, and DeepSeek-V3.

RobotSearch is an automatic literature classification tool specifically designed for identifying RCTs. It employs machine-learning algorithms trained on a large dataset of RCTs from the Cochrane Crowd. This tool, developed by Marshall et al., is freely available to the public [24].

In contrast, ChatGPT 4.0, Claude 3.5, Gemini 1.5, and DeepSeek-V3 are large language models (LLMs) that were not specifically designed for RCT classification. However, these LLMs have demonstrated promising performance in general text-classification tasks [10, 16, 25].

2.4 Prompt engineering approach

The prompt-engineering process consisted of three key steps. First, primary prompts were carefully developed and refined for the literature-screening task, with the assistance of LLMs to improve the initial designs. For example, a prompt such as "Please design the best prompt for me based on this prompt: ..." was used during this step [26]. Second, iterative testing was conducted to further optimize the prompts. Finally, the refined prompts were applied to ChatGPT, Claude, Gemini, and DeepSeek to perform the literature-screening tasks. Details are presented in Table 1.

Table 1: Prompt used for literature screening

```
prompt=(
"Determine whether the provided literature (title, abstract)
represents a randomized controlled trial (RCT).\n"
f"Please make a judgment based on the following paper title
and abstract:\nTitle: {title} \nAbstract: {abstract}\n"
"1. Review the title, abstract, and content of the PDF document.\n"
"2. Determine if the study involves random assignment of
participants to different interventions.\n"
"3. Consider key indicators of RCTs such as \"randomized\",
\"controlled\", \"trial\", \"random allocation\", and
\"random assignment\".\n"
"4. If the study explicitly states it is an RCT or meets the
criteria of an RCT, classify it as \"yes\".\n"
"5. If the study does not meet the criteria or lacks sufficient
information to determine, classify it as \"no\".\n"
"output_format:\n"
"{\n"
"\"result\": \"yes\" or \"no\"\n"
"}\n"
"Output the json data only, with no additional text.")
```

2.5 Outcomes

For automated literature-screening tools, the study assumed that the most practically valuable tool should possess two key properties: (1) a near-zero false negative fraction (FNF), defined as

the proportion of RCTs that are incorrectly excluded; and (2) a screening speed significantly faster than that of human reviewers.

Accordingly, the primary outcome of this study was the false negative fraction of the tools. Secondary outcomes included the total time required for screening both the RCT and non-RCT groups and the redundancy number needed to screen (RNNS). The RNNS was defined as the number of studies in the RCT group that would still require manual review after the automated screening process, specifically those incorrectly retained by the tool.

2.6 Statistical analysis

We compared the false negative fraction (FNF) and its 95% confidence interval (CI) for each tool in the RCT group, and the false positive fraction (FPF) and its 95% CI in the non-RCT group. A lower FNF indicates a reduced likelihood of excluding studies that should be included, while a lower FPF reflects a reduced likelihood of including studies that should be excluded, or fewer studies requiring further manual screening.

We calculated the positive likelihood ratio,

$$\text{PLR} = \frac{\text{Sensitivity}}{1 - \text{Specificity}},$$

to evaluate the diagnostic performance of each AI tool. Youden's Index,

$$J = \text{Sensitivity} + \text{Specificity} - 1,$$

provided additional insight into overall test performance and allowed comparison of the effectiveness of the different screening tools.

The area under the curve (AUC) was not estimated because, in the context of literature screening for evidence synthesis, minimizing the false negative fraction, $1 - \text{Sensitivity}$, was prioritized over minimizing the false positive fraction, $1 - \text{Specificity}$.

Percentage differences between the AI tools and the time required for screening were also calculated, with a significance level of $\alpha = 0.05$. Risk difference,

$$\text{RD} = \text{Risk in Exposed Group} - \text{Risk in Unexposed Group},$$

and risk ratio,

$$\text{RR} = \frac{\text{Risk in Exposed Group}}{\text{Risk in Unexposed Group}},$$

were calculated to complement the analysis.

The number needed to screen (NNS), calculated as

$$\text{NNS} = \frac{1}{J},$$

provided a practical metric for understanding the efficiency of each screening tool. All statistical analyses were performed using Stata SE 16.0 (StataCorp LLC, College Station, TX).

3 Results

Table 2 outlines the baseline characteristics of the included studies. Of the 1,000 publications, 500 (50.0%) were randomized controlled trials (RCTs), while the remaining 500 (50.0%) were classified as non-RCTs. Approximately one-third of the trials (36.0%) were published after 2019.

Table 2: Baseline characteristics of the included studies

Characteristic	No. of studies (n=1000)
Type of design	
RCTs	500 (50.0%)
Others	500 (50.0%)
Year of retraction	
Before 2010	364 (36.4%)
2011–2019	487 (48.7%)
After 2019	149 (14.9%)
Year of publication	
Before 2010	93 (9.3%)
2011–2019	547 (54.7%)
After 2019	360 (36.0%)

3.1 Performance of automatic screening tools

3.1.1 False negative fraction

Within the RCT group, RobotSearch demonstrated the lowest false negative fraction (FNF) at 6.4% (95% CI: 4.6%–8.9%) among the evaluated AI tools. In contrast, Gemini exhibited the highest FNF at 13.0% (95% CI: 10.3%–16.3%). Figure 1 summarizes the performance of the evaluated tools.

Significant differences in FNF were observed between Gemini and several other AI tools, including ChatGPT, DeepSeek, and RobotSearch. A significant difference was also observed between Claude and RobotSearch. No statistically significant differences in FNF were found among ChatGPT, Claude, and DeepSeek (Figure 2).

3.1.2 Time used for screening

The time required for literature screening was recorded using the automatic timers of the large language models. The total screening times for the RCT and non-RCT groups were 22.0 minutes for ChatGPT, 100.0 minutes for Claude, 20.0 minutes for Gemini, and 43.5 minutes for DeepSeek.

On average, the mean screening time per article was 1.3 s for ChatGPT, 6.0 s for Claude, 1.2 s for Gemini, and 2.6 s for DeepSeek.

3.1.3 False positive fraction and redundancy number needed to screen

Within the non-RCT group, the false positive fraction (FPF) for the four large language models was below 4.0%. Gemini achieved the lowest FPF at 2.8% (95% CI: 1.7%–4.7%), whereas RobotSearch exhibited the highest FPF at 22.2% (95% CI: 18.8%–26.1%).

The redundancy numbers needed to screen (RNNS) for ChatGPT, Claude, Gemini, and DeepSeek were 16, 14, 19, and 17, respectively. In comparison, RobotSearch required screening of 111 redundant articles.

RobotSearch exhibited significantly higher FPF than all four LLMs. However, no statistically significant differences in FPF were observed among ChatGPT, Claude, Gemini, and DeepSeek (Figure 2).

3.1.4 Diagnostic performance evaluation

A comprehensive analysis revealed that ChatGPT demonstrated the strongest overall literature-screening performance, with a positive likelihood ratio (PLR) of 28.812, Youden's Index of 0.89,

number needed to screen (NNS) of 1.123, and risk difference (RD) and risk ratio (RR) values of 0.89 and 28.813, respectively.

Claude and DeepSeek showed diagnostic performance comparable to ChatGPT, whereas Gemini demonstrated somewhat lower overall performance. RobotSearch showed substantially weaker performance than the four LLMs, with a PLR of 4.216, Youden's Index of 0.714, and NNS of 1.4. Overall, all four large language models outperformed RobotSearch with respect to both false negative fraction and false positive fraction (see Supplementary Table S1).

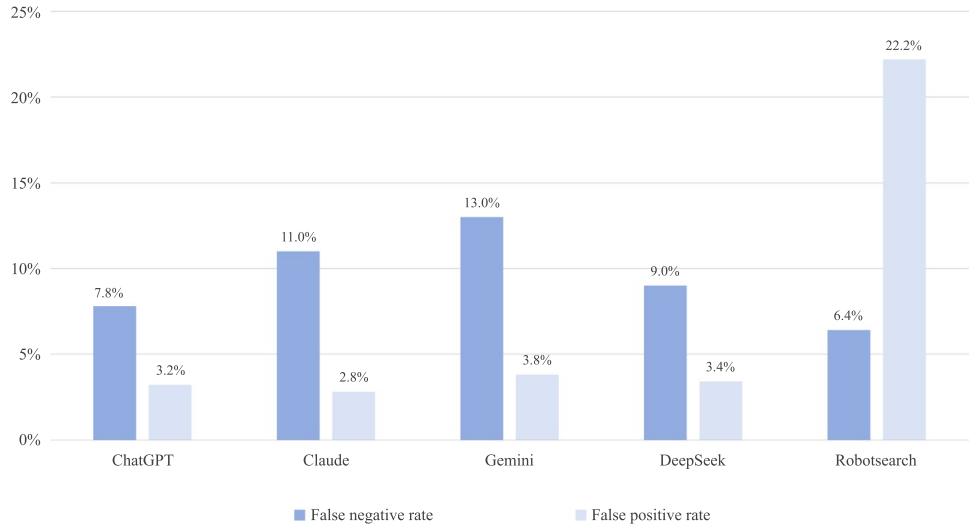

Figure 1: Performance of the AI tools for literature screening.

False negative rate	ChatGPT	Claude	Gemini	DeepSeek	RobotSearch	**False positive rate**
ChatGPT		0.4%(-1.7%,2.5%)	-0.6%(-2.9%,1.7%)	-0.2%(-2.4%,2.0%)	-19.0%(-23.0%,-15.0%)*	ChatGPT
Claude	-3.2%(-6.8%,0.4%)		-1.0%(-3.2%,1.2%)	-0.6%(-2.7%,1.5%)	-19.4%(-23.3%,-15.5%)*	Claude
Gemini	-5.2%(-9.0%,-1.4%)*	-2.0%(-6.0%,2.0%)		0.4%(-1.9%,2.7%)	-18.4%(-22.4%,-14.4%)*	Gemini
DeepSeek	-1.2%(-4.6%,2.2%)	2%(-1.7%,5.7%)	4.0%(0.1%,7.9%)*		-18.8%(-22.8%,-14.8%)*	DeepSeek
RobotSearch	1.4%(-1.8%,4.6%)	4.6%(1.1%,8.1%)*	6.6%(3.0%,10.2%)*	2.6%(-0.7%,5.9%)		RobotSearch
	ChatGPT	Claude	Gemini	DeepSeek	RobotSearch	

*$P < 0.05$

Figure 2: Percentage difference in the performance of the AI tools.

4 Discussion

In this investigation, we assessed the performance of five automated literature-screening tools: ChatGPT, Claude, Gemini, DeepSeek, and RobotSearch.

Regarding screening efficiency, ChatGPT, Gemini, and DeepSeek demonstrated a mean processing time of less than 3 s per article, whereas Claude required approximately 6 s per article. Consistent with prior studies, these tools exhibit considerable potential for improving the

efficiency of literature screening. Human screening of the same dataset required approximately two weeks, highlighting the substantial time savings offered by AI-assisted screening. Nevertheless, the results indicate that AI tools cannot yet function independently without human supervision. Future work should investigate strategies that integrate human expertise with AI-driven screening to further improve screening performance.

Among the evaluated AI tools, RobotSearch exhibited the lowest false negative fraction (FNF) at 6.4%, whereas Gemini had the highest FNF at 13.0%. Within the non-RCT group, the false positive fraction (FPF) for the four large language models ranged from 2.8% to 3.8%, substantially lower than that observed for RobotSearch. Minimizing the FNF is particularly important because it reduces the probability of excluding studies that should be included in evidence synthesis. Although the AI tools demonstrated relatively low FNF values compared with some previous studies, their performance remained inferior to that of human reviewers. Compared with earlier reports, ChatGPT achieved a comparatively low FNF, whereas Claude and Gemini exceeded 10%, with Gemini demonstrating the highest value. The authors suggest that Gemini's comparatively poorer performance may be related to limitations in processing complex medical texts.

The study also found that large language models consistently achieved lower FPF than traditional automated screening tools such as RobotSearch. All four evaluated LLMs produced FPF values below 4%, requiring fewer than twenty articles for additional manual review. These findings highlight the potential of LLMs to substantially reduce manual screening workload. Compared with traditional machine-learning tools, LLMs benefit from improved contextual understanding and semantic representation, enabling more accurate identification of relevant literature and improved screening efficiency.

Diagnostic performance metrics further demonstrated differences among the evaluated models. ChatGPT achieved the highest Youden's Index (0.89), indicating the best balance between sensitivity and specificity. It also required the least manual re-screening according to the Number Needed to Screen (NNS). Claude and DeepSeek demonstrated comparable performance, whereas RobotSearch consistently performed worse across the evaluated diagnostic measures. Overall, the evaluated LLMs demonstrated considerable potential for improving literature-screening workflows.

The authors further noted that DeepSeek-V3, which has recently attracted attention because of its upgraded capabilities, performed similarly to ChatGPT in this evaluation. Consequently, both ChatGPT and DeepSeek were considered robust tools for literature screening. Nevertheless, the authors emphasized that additional validation using larger and more diverse datasets is required.

To the best of the authors' knowledge, this study represents the first comparison employing multiple large language models for fully automated literature screening. The enhanced data-processing capabilities of LLMs offer several practical advantages. They can rapidly process large volumes of literature, reduce errors associated with human fatigue and subjective bias, lower research costs, and continuously adapt to evolving research trends. Despite these advantages, the authors concluded that LLMs should currently be regarded as supplementary tools rather than replacements for human reviewers. Continued model development together with human oversight remains essential for reliable evidence synthesis.

Limitations

Several limitations were acknowledged. First, the literature cohort consisted of retracted publications available through the Retraction Watch database up to April 26, 2023, and did not include databases covering additional disease types, which may influence generalizability. Second,

because of the inherent variability of LLM outputs, the consistency of repeated runs using identical prompts was not evaluated. Finally, the study focused exclusively on distinguishing randomized controlled trials from other study types. The effectiveness of the evaluated approach for screening other forms of literature remains to be established in future work.

References

[1] J. Gurevitch, J. Koricheva, S. Nakagawa, and G. Stewart. Meta-analysis and the science of research synthesis. *Nature*, 555(7695):175–182, 2018. doi: 10.1038/nature25753. URL https://doi.org/10.1038/nature25753.

[2] C. Xu, T. Yu, L. Furuya-Kanamori, L. Lin, L. Zorzela, and X. Zhou. Validity of data extraction in evidence synthesis practice of adverse events: reproducibility study. *BMJ*, 377:e069155, 2022. doi: 10.1136/bmj-2021-069155. URL https://doi.org/10.1136/bmj-2021-069155.

[3] E. Aromataris, R. Fernandez, C. M. Godfrey, C. Holly, H. Khalil, and P. Tungpunkom. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. *International Journal of Evidence-Based Healthcare*, 13(3): 132–140, 2015. doi: 10.1097/XEB.0000000000000055. URL https://doi.org/10.1097/XEB.0000000000000055.

[4] M. Cumpston, T. Li, M. J. Page, J. Chandler, V. A. Welch, J. P. T. Higgins, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. *Cochrane Database of Systematic Reviews*, 10(10): ED000142, 2019.

[5] M. J. Page, D. Moher, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372:n160, 2021. doi: 10.1136/bmj.n160. URL https://doi.org/10.1136/bmj.n160.

[6] R. Borah, A. W. Brown, P. L. Capers, and K. A. Kaiser. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*, 7(2):e012545, 2017. doi: 10.1136/bmjopen-2016-012545. URL https://doi.org/10.1136/bmjopen-2016-012545.

[7] M. Michelson and K. Reuter. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary Clinical Trials Communications*, 16:100443, 2019. doi: 10.1016/j.conctc.2019.100443. URL https://doi.org/10.1016/j.conctc.2019.100443.

[8] R. F. Soll, C. Ovelman, and W. McGuire. The future of Cochrane Neonatal. *Early Human Development*, 150:105191, 2020. doi: 10.1016/j.earlhumdev.2020.105191. URL https://doi.org/10.1016/j.earlhumdev.2020.105191.

[9] Y. Lin, J. Li, H. Xiao, L. Zheng, Y. Xiao, H. Song, et al. Automatic literature screening using the PAJO deep-learning model for clinical practice guidelines. *BMC Medical Informatics and Decision Making*, 23(1):247, 2023. doi: 10.1186/s12911-023-02328-8. URL https://doi.org/10.1186/s12911-023-02328-8.

[10] F. M. Delgado-Chaves, M. J. Jennings, A. Atalaia, J. Wolff, R. Horvath, Z. M. Mamdouh, et al. Transforming literature screening: the emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences of the United States of America*, 122(2):e2411962122, 2025. doi: 10.1073/pnas.2411962122. URL https://doi.org/10.1073/pnas.2411962122.

[11] J. Kamtchum-Tatuene and J. G. Zafack. Keeping up with the medical literature: why, how, and when? *Stroke*, 52(11):e746–e748, 2021. doi: 10.1161/STROKEAHA.121.036141. URL https://doi.org/10.1161/STROKEAHA.121.036141.

[12] T. Li, J. P. T. Higgins, and J. J. Deeks. Chapter 5: Collecting data. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch, editors, *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane, version 6.3, updated february 2022 edition, 2022.

[13] S. Waffenschmidt, M. Knelangen, W. Sieben, S. Bühn, and D. Pieper. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Medical Research Methodology*, 19(1):132, 2019. doi: 10.1186/s12874-019-0782-0. URL https://doi.org/10.1186/s12874-019-0782-0.

[14] T. Tsubota, D. Bollegala, Y. Zhao, Y. Jin, and T. Kozu. Improvement of intervention information detection for automated clinical literature screening during systematic review. *Journal of Biomedical Informatics*, 134:104185, 2022. doi: 10.1016/j.jbi.2022.104185. URL https://doi.org/10.1016/j.jbi.2022.104185.

[15] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023. doi: 10.1038/s41591-023-02448-8. URL https://doi.org/10.1038/s41591-023-02448-8.

[16] Q. Khraisha, S. Put, J. Kappenberg, A. Warraitch, and K. Hadfield. Can large language models replace humans in systematic reviews? evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*, 15(4):616–626, 2024. doi: 10.1002/jrsm.1715. URL https://doi.org/10.1002/jrsm.1715.

[17] A. Blaizot, S. K. Veettil, P. Saidoung, C. F. Moreno-Garcia, N. Wiratunga, M. Aceves-Martins, et al. Using artificial intelligence methods for systematic review in health sciences: a systematic review. *Research Synthesis Methods*, 13(3):353–362, 2022. doi: 10.1002/jrsm.1553. URL https://doi.org/10.1002/jrsm.1553.

[18] M. Goldkuhle, M. Dimaki, G. Gartlehner, I. Monsef, P. Dahm, J. P. Glossmann, et al. Nivolumab for adults with Hodgkin's lymphoma (a rapid review using the software RobotReviewer). *Cochrane Database of Systematic Reviews*, 7(7):CD012556, 2018.

[19] M. M. Kebede, C. Le Cornet, and R. T. Fortner. In-depth evaluation of machine learning methods for semi-automating article screening in a systematic review of mechanistic literature. *Research Synthesis Methods*, 14(2):156–172, 2023. doi: 10.1002/jrsm.1589. URL https://doi.org/10.1002/jrsm.1589.

[20] A. O. D. Santos, E. S. da Silva, L. M. Couto, G. V. L. Reis, and V. S. Belo. The use of artificial intelligence for automating or semi-automating biomedical literature analyses: a scoping review. *Journal of Biomedical Informatics*, 142:104389, 2023. doi: 10.1016/j.jbi.2023.104389. URL https://doi.org/10.1016/j.jbi.2023.104389.

[21] A. Valizadeh, M. Moassefi, A. Nakhostin-Ansari, S. H. Hosseini Asl, M. Saghab Torbati, R. Aghajani, et al. Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. *BMC Medical Research Methodology*, 22(1):160, 2022. doi: 10.1186/s12874-022-01631-8. URL https://doi.org/10.1186/s12874-022-01631-8.

[22] K. Dwan, T. Li, D. G. Altman, and D. Elbourne. CONSORT 2010 statement: extension to randomised crossover trials. *BMJ*, 366:l4378, 2019. doi: 10.1136/bmj.l4378. URL https://doi.org/10.1136/bmj.l4378.

[23] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. M. Irwig, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*, 326(7379):41–44, 2003. doi: 10.1136/bmj.326.7379.41. URL https://doi.org/10.1136/bmj.326.7379.41.

[24] I. J. Marshall, A. Noel-Storr, J. Kuiper, J. Thomas, and B. C. Wallace. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. *Research Synthesis Methods*, 9(4):602–614, 2018. doi: 10.1002/jrsm.1287. URL https://doi.org/10.1002/jrsm.1287.

[25] E. Guo, M. Gupta, J. Deng, Y. J. Park, M. Paget, and C. Naugler. Automated paper screening for clinical reviews using large language models: data analysis study. *Journal of Medical Internet Research*, 26:e48996, 2024. doi: 10.2196/48996. URL https://doi.org/10.2196/48996.

[26] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198, 2023. doi: 10.1371/journal.pdig.0000198. URL https://doi.org/10.1371/journal.pdig.0000198.