Original Paper

Accuracy of Large Language Models for Literature Screening in Thoracic Surgery: Diagnostic Study

Zhang-Yi Dai, MD; Fu-Qiang Wang, MD; Cheng Shen, MDS; Yan-Li Ji, MMSc; Zhi-Yang Li, MD; Yun Wang, PhD; Qiang Pu, MD, PhD

Department of Thoracic Surgery, West China Hospital of Sichuan University, Chengdu, China

Corresponding Author:
Qiang Pu, MD, PhD
Department of Thoracic Surgery
West China Hospital of Sichuan University
No.37, Guoxue Alley
Chengdu, 610041
China
Phone: 86 18980606738
Email: puqiang100@163.com

Abstract

Background: Systematic reviews and meta-analyses rely on labor-intensive literature screening. While machine learning offers potential automation, its accuracy remains suboptimal. This raises the question of whether emerging large language models (LLMs) can provide a more accurate and efficient approach.

Objective: This paper evaluates the sensitivity, specificity, and summary receiver operating characteristic (SROC) curve of LLM-assisted literature screening.

Methods: We conducted a diagnostic study comparing the accuracy of LLM-assisted screening versus manual literature screening across 6 thoracic surgery meta-analyses. Manual screening by 2 investigators served as the reference standard. LLM-assisted screening was performed using ChatGPT-4o (OpenAI) and Claude-3.5 (Anthropic) sonnet, with discrepancies resolved by Gemini-1.5 pro (Google). In addition, 2 open-source, machine learning–based screening tools, ASReview (Utrecht University) and Abstrackr (Center for Evidence Synthesis in Health, Brown University School of Public Health), were also evaluated. We calculated sensitivity, specificity, and 95% CIs for the title and abstract, as well as full-text screening, generating pooled estimates and SROC curves. LLM prompts were revised based on a post hoc error analysis.

Results: LLM-assisted full-text screening demonstrated high pooled sensitivity (0.87, 95% CI 0.77-0.99) and specificity (0.96, 95% CI 0.91-0.98), with the area under the curve (AUC) of 0.96 (95% CI 0.94-0.97). Title and abstract screening achieved a pooled sensitivity of 0.73 (95% CI 0.57-0.85) and specificity of 0.99 (95% CI 0.97-0.99), with an AUC of 0.97 (95% CI 0.96-0.99). Post hoc revisions improved sensitivity to 0.98 (95% CI 0.74-1.00) while maintaining high specificity (0.98, 95% CI 0.94-0.99). In comparison, the pooled sensitivity and specificity of ASReview tool-assisted screening were 0.58 (95% CI 0.53-0.64) and 0.97 (95% CI 0.91-0.99), respectively, with an AUC of 0.66 (95% CI 0.62-0.70). The pooled sensitivity and specificity of Abstrackr tool-assisted screening were 0.48 (95% CI 0.35-0.62) and 0.96 (95% CI 0.88-0.99), respectively, with an AUC of 0.78 (95% CI 0.74-0.82). A post hoc meta-analysis revealed comparable effect sizes between LLM-assisted and conventional screening.

Conclusions: LLMs hold significant potential for streamlining literature screening in systematic reviews, reducing workload without sacrificing quality. Importantly, LLMs outperformed traditional machine learning-based tools (ASReview and Abstrackr) in both sensitivity and AUC values, suggesting that LLMs offer a more accurate and efficient approach to literature screening.

(J Med Internet Res 2025;27:e67488) doi: 10.2196/67488

KEYWORDS

accuracy; large language models; meta-analysis; literature screening; thoracic surgery

Introduction

The development of clinical practice guidelines necessitates a comprehensive and systematic synthesis of current research evidence [1]. Evidence-based medicine frequently relies on systematic reviews and meta-analyses, which aggregate findings from studies on a specific topic [2-4]. This process traditionally involves extensive effort in identifying and retrieving relevant literature [5,6]. While machine learning has shown promise in streamlining literature retrieval, its accuracy often falls short of desired standards [7-9]. Therefore, further research is needed to develop a more precise screening method.

Recently, large language models (LLMs) powered by natural language processing have demonstrated remarkable capabilities in various domains, including language comprehension, image and video generation, and data analysis [8-13]. Previous studies have suggested the potential of LLMs for literature screening [9,14,15]. However, their accuracy in the screening process for meta-analyses remains unclear.

We hypothesized that LLM-assisted literature screening could achieve accuracy comparable with manual screening. To test this hypothesis, we designed a diagnostic trial using conventional manual screening as the reference standard to evaluate the accuracy of LLM-assisted literature screening.

Methods

Study Design

This prospective diagnostic study aimed to assess the validity of LLMs for assisting with literature screening during meta-analysis. Conventional literature manual screening served as the reference standard. This diagnostic study was performed according to the Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidelines and the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist.

Before the study commenced, we defined the research topic as the comparison between sublobar resection and lobectomy in thoracic surgery, a topic of ongoing debate within the field. A total of 6 relevant published meta-analyses were identified [16-21], and the search strategy and terms were subsequently redesigned for a new round of literature retrieval and screening. To ensure a comprehensive yet nonredundant literature base, the search strategies and inclusion timeframes outlined within each meta-analysis were replicated. Duplicate studies were subsequently identified and removed using Rayyan [22], a web-based literature management tool.

Conventional Literature Screening

Following identification, 2 independent investigators (Lei Peng and Xing-Yu Liu) screened the titles and abstracts of retrieved studies for inclusion based on predefined criteria (detailed in Table S1 in Multimedia Appendix 1). Discrepancies were resolved through adjudication by a third investigator (Xu-Yang Wang). Full-text papers of potentially eligible studies were subsequently reviewed by the same 2 independent investigators (Lei Peng and Xing-Yu Liu) against the same inclusion criteria, and any discrepancies were again resolved through adjudication by the third investigator (Xu-Yang Wang). This conventional literature screening established the reference standard for comparison. Investigators involved in the conventional screening were excluded from participation in the LLM-assisted screening and subsequent analyses.

LLM-Assisted Literature Screening

For the LLM-assisted literature screening, a 5-column table (author, publication year, journal, title, and abstract) was compiled from the deduplicated literature. Following established prompt engineering guidelines [23], specific prompts were developed to facilitate automated screening using Python (version 3.9.0; Python Software Foundation). These prompts, structured to output results in a tabular format, instructed the LLM to perform screening based on the Population, Intervention, Control, Outcome, and Study design (PICOS) framework criteria defined for each topic study (detailed in Table S1 in Multimedia Appendix 1). An example prompt is provided in Figure S1 in Multimedia Appendix 1.

In addition, 2 LLMs, ChatGPT-4o (OpenAI) and Claude-3.5 sonnet (Anthropic), were used as independent reviewers to independently screen titles and abstracts. Study selection was based on the predefined inclusion and exclusion criteria. Discrepancies in study selection between the 2 LLMs were resolved by Gemini-1.5 pro (Google). Full-text papers underwent an identical screening process. The literature screening process assisted by LLM was conducted and supervised by 2 reviewers (Z-YD and F-QW). The detailed prompts used in LLM-assisted literature screening are provided in section S1 in Multimedia Appendix 2.

We then evaluated the performance of 2 open-source, machine learning–based screening tools, ASReview [24,25] and Abstrackr [26,27], for title and abstract screening. We compared their results against conventional manual screening methods, which served as the reference standard, to assess the accuracy of LLM-assisted literature screening. The detailed methodology is documented in section S2 in Multimedia Appendix 2.

Statistical Analysis

The accuracy of both LLM-assisted literature screening and 2 semiautomated, machine learning–based screening tools were evaluated in each topical study using sensitivity, specificity, and their corresponding 95% CIs. The primary analysis focused on the sensitivity and specificity of LLM-assisted literature screening assessed after full-text review. The secondary analysis focused on the sensitivity and specificity of LLM-assisted screening at the title and abstract review stage. In addition, meta-analysis techniques were used to calculate pooled sensitivity and specificity, along with the summary receiver operating characteristic (SROC) curve. This provided overall results for the primary, secondary, and post hoc analyses. Heterogeneity across topic studies was assessed by calculating the inconsistency value (I^2) using the chi-square test. A random-effects model was used to pool sensitivity and specificity if I^2 exceeded 50% or if the P value was less than .05. Meta-analyses were conducted using Stata (version 15.0;

StataCorp), while other statistical analyses were performed using GraphPad Prism (version 8.0.1; GraphPad Prism, Inc).

A post hoc analysis (conducted by YW and QP) involved reviewing papers of false-negative classifications to identify the sources of LLM errors during screening (the comprehensive explanations for the occurrence of false-negative classifications are provided in Table S2 in Multimedia Appendix 1). Subsequently, the literature screening prompts were refined by incorporating a chain-of-thought prompting strategy [28] based on the identified error patterns (conducted by CS). In addition, 3 iterations of screening querying were then performed with revised prompts to optimize the model's validity. A study was considered eligible if the LLM classified it as eligible during any of the 3 iterations.

To further assess the robustness of LLM-assisted screening and account for potential variations, a separate post hoc meta-analysis was conducted (Y-LJ and Z-YL). This analysis compared the pooled effect sizes derived from LLM-assisted screening (including only true positives) with those from conventional screening (including both true positives and false negatives) for each topic study.

Ethical Considerations

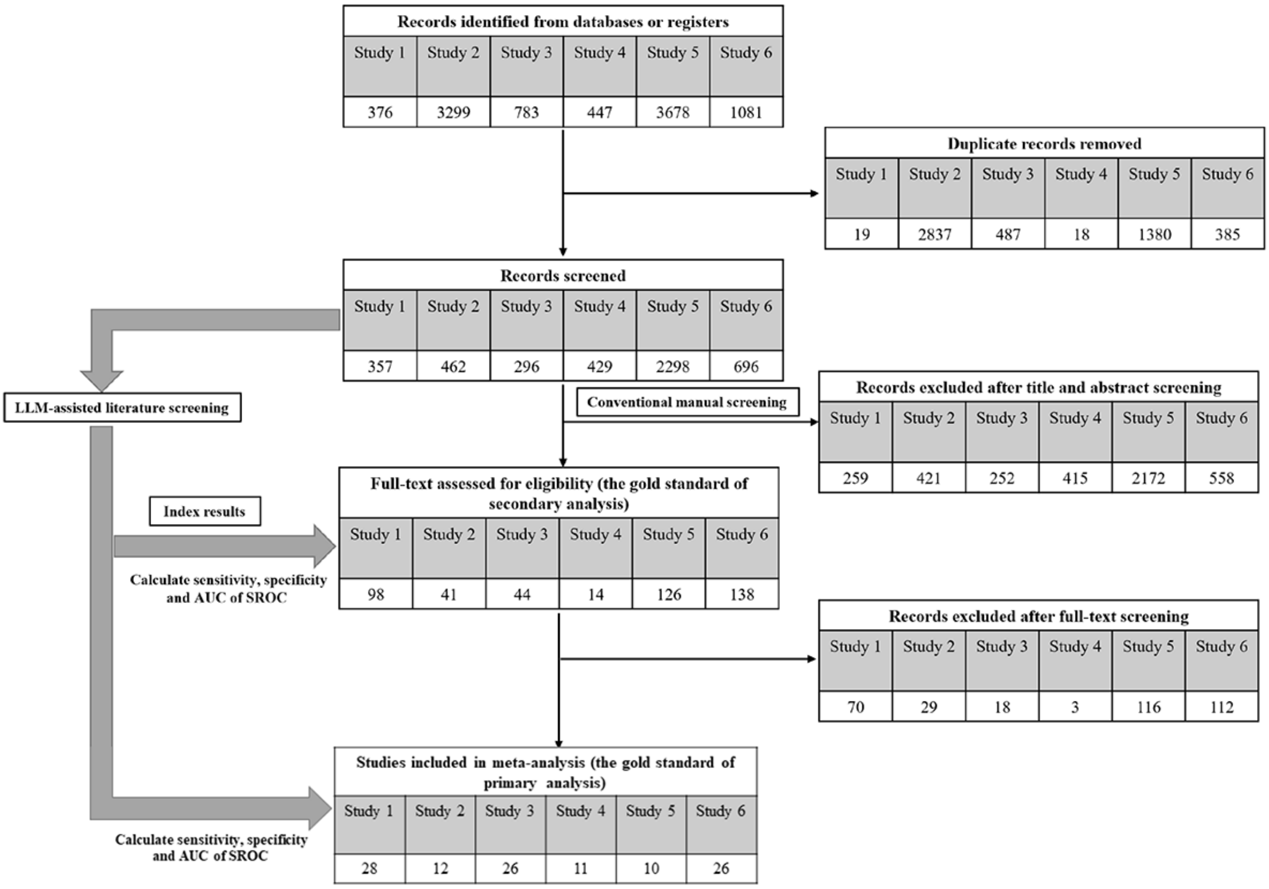
The study did not involve human participants or biological specimens. As such, the Biomedical Ethics Committee of West China Hospital, Sichuan University determined that this research was eligible for exemption from ethical review (reference number: 2024−1177). This decision aligns with institutional and local policies that exempt studies not involving human subjects or biological materials from requiring formal ethics board approval.

Results

The Results of Conventional Manual and LLM-Assisted Literature Screening

Following deduplication in the conventional literature screening process, the initial search yielded 357, 462, 296, 429, 2,298, and 696 papers for studies 1-6, respectively. Title and abstract screening resulted in 98, 41, 44, 14, 126, and 138 papers selected for full-text review in the corresponding studies. Ultimately, 28, 12, 26, 11, 10, and 26 papers from studies 1 to 6, respectively, met the inclusion criteria and were incorporated into the final meta-analysis (Figure 1). The results of LLM-assisted literature screening are listed and described in Table S3 in Multimedia Appendix 1 and section S3 in Multimedia Appendix 2.

Figure 1. Flow diagram of meta-analysis using conventional manual literature screening and large language model (LLM)–assisted screening. AUC: area under the curve; SROC: summary receiver operating characteristic curve.



Records identified from databases or registers

Study 1	Study 2	Study 3	Study 4	Study 5	Study 6
376	3299	783	447	3678	1081

Duplicate records removed

Study 1	Study 2	Study 3	Study 4	Study 5	Study 6
19	2837	487	18	1380	385

Records screened

Study 1	Study 2	Study 3	Study 4	Study 5	Study 6
357	462	296	429	2298	696

Records excluded after title and abstract screening

Study 1	Study 2	Study 3	Study 4	Study 5	Study 6
259	421	252	415	2172	558

Full-text assessed for eligibility (the gold standard of secondary analysis)

Study 1	Study 2	Study 3	Study 4	Study 5	Study 6
98	41	44	14	126	138

Records excluded after full-text screening

Study 1	Study 2	Study 3	Study 4	Study 5	Study 6
70	29	18	3	116	112

Studies included in meta-analysis (the gold standard of primary analysis)

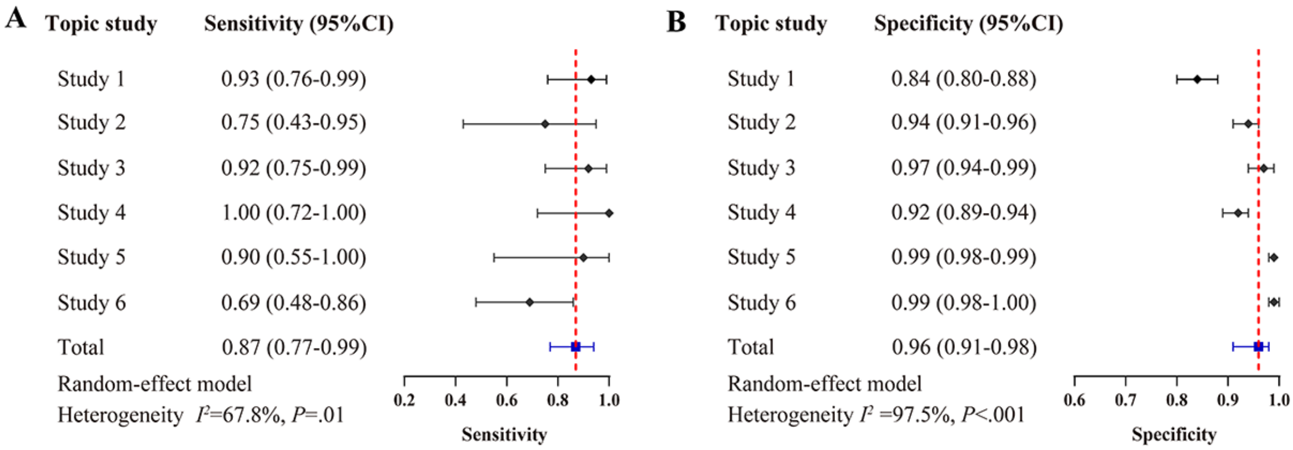
Study 1	Study 2	Study 3	Study 4	Study 5	Study 6
28	12	26	11	10	26

LLM-assisted literature screening
Conventional manual screening
Index results
Calculate sensitivity, specificity and AUC of SROC
Calculate sensitivity, specificity and AUC of SROC

Primary Analysis of LLM-Assisted Literature Screening

In the LLM-assisted literature screening process, a total of 26, 9, 24, 11, 9, and 18 papers from studies 1 to 6, respectively, were included in the final meta-analysis (Table S3 in Multimedia Appendix 1). The primary analysis revealed the sensitivity and specificity of the LLM-assisted screening for studies 1-6 as follows: 0.93 (95% CI 0.76-0.99) and 0.84 (95% CI 0.80-0.88), 0.75 (95% CI 0.43-0.95) and 0.94 (95% CI 0.91-0.96), 0.92 (95% CI 0.75-0.99) and 0.97 (95% CI 0.94-0.99), 1.00 (95% CI 0.72-1.00) and 0.92 (95% CI 0.89-0.94), 0.90 (95% CI 0.55-1.00) and 0.99 (95% CI 0.98-0.99), and 0.69 (95% CI 0.48-0.86) and 0.99 (95% CI 0.98-1.00), respectively (Figure 2). Meta-analysis of 6 topic studies revealed that LLM-assisted screening demonstrated a high discriminative ability, with SROC curve analysis yielding an area under the curve (AUC) of 0.96 (95% CI 0.94-0.97); see Figure S2 in Multimedia Appendix 1. Furthermore, the pooled sensitivity and specificity were 0.87 (95% CI 0.77-0.99) and 0.96 (95% CI 0.91-0.98), respectively (Figure 2). The counts of true positives, false negatives, false positives, and true negatives of primary analysis are detailed in Table S3 in Multimedia Appendix 1.

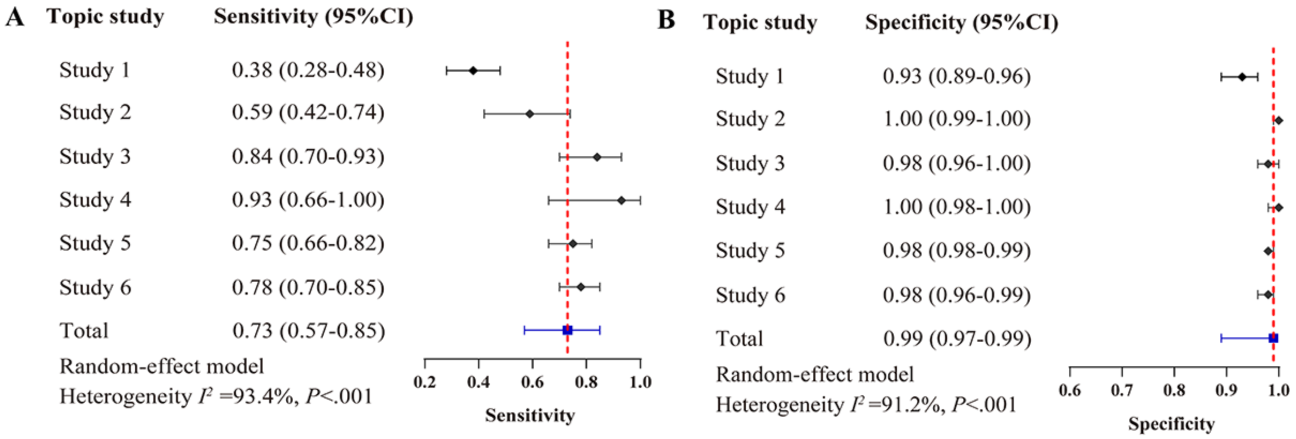
Figure 2. Sensitivity and specificity of large language model (LLM)–assisted literature screening in the primary analysis.



Secondary Analysis of LLM-Assisted Screening

The pooled sensitivity and specificity across these 6 studies were 0.73 (95% CI 0.57-0.85) and 0.99 (95% CI 0.97-0.99), respectively (Figure 3 and section S4 in Multimedia Appendix 2). The SROC analysis yielded an AUC of 0.97 (95% CI 0.96-0.99); see Figure S2 in Multimedia Appendix 1. The counts of true positives, false negatives, false positives, and true negatives of secondary analysis are detailed in Table S3 in Multimedia Appendix 1.

Figure 3. Sensitivity and specificity of large language model (LLM)–assisted literature screening in the secondary analysis.

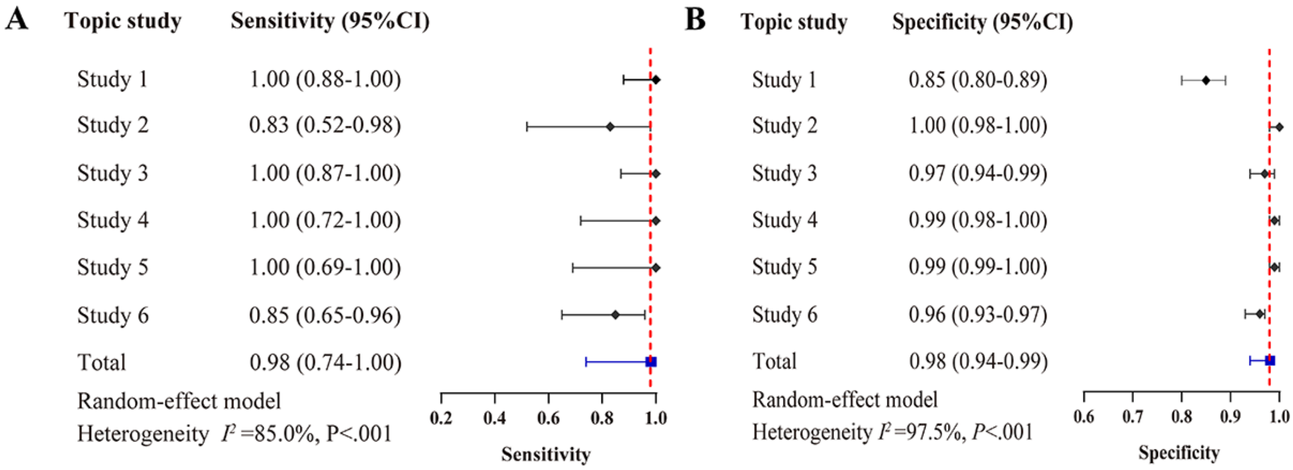


The pooled sensitivity and specificity of ASReview tool-assisted screening were 0.58 (95% CI 0.53-0.64) and 0.97 (95% CI 0.91-0.99), respectively (Figure S3 in Multimedia Appendix 1). The pooled sensitivity and specificity of Abstrackr tool-assisted screening were 0.48 (95% CI 0.35-0.62) and 0.96 (95% CI 0.88-0.99), respectively (Figure S4 in Multimedia Appendix 1). The SROC analysis yielded AUC values of 0.66 (95% CI 0.62-0.70) and 0.78 (95% CI 0.74-0.82), respectively (Figure S5 in Multimedia Appendix 1). The corresponding counts of true positives, false negatives, false positives, and true negatives of the title and abstract screening phase are detailed in Table S4 in Multimedia Appendix 1.

Post Hoc Analysis of Revised Prompts

A post hoc analysis was conducted using a revised prompt (Figure S6 in Multimedia Appendix 1) and incorporating a chain-of-thought strategy (Table S2 in Multimedia Appendix 1 and section S5 in Multimedia Appendix 2). The sensitivity and specificity of LLM-assisted screening in the primary analysis are described in section S6 in Multimedia Appendix 2. The overall sensitivity and specificity across these 6 studies in the primary analysis were 0.98 (95% CI 0.74-1.00) and 0.98 (95% CI 0.94-0.99), respectively (Figure 4). The SROC analysis yielded an AUC of 1.00 (95% CI 0.99-1.00; Figure S7 in Multimedia Appendix 1). The counts of true positives, false negatives, false positives, and true negatives for the post hoc analysis are detailed in Table S3 in Multimedia Appendix 1. The sensitivity, specificity, and AUC of LLM-assisted screening in the secondary analysis are presented and described in Figures S7-S8 in Multimedia Appendix 1 and section S7 in Multimedia Appendix 2.

Figure 4. Sensitivity and specificity of large language model (LLM)–assisted literature screening: primary analysis with revised prompt (post hoc).



Post Hoc Meta-Analysis of Index Results of LLM-Assisted Screening

This post hoc meta-analysis compared pooled effect sizes from studies 1, 2, 3, 5, and 6 (Table S5 in Multimedia Appendix 1 and section S8 in Multimedia Appendix 2) for LLM-assisted screening versus conventional screening. The meta-analysis revealed comparable results between the two methods. Furthermore, the false-negative papers did not substantially affect the overall conclusions of the corresponding topic studies (Figures S9-S17 in Multimedia Appendix 1).

Discussion

Principal Findings

Our study addresses a critical challenge in the development of clinical practice guidelines—the labor-intensive and time-consuming nature of literature screening in meta-analyses [2]. Traditionally, this process relies heavily on manual efforts to ensure the inclusion of high-quality evidence from randomized controlled trials and cohort studies [2,3,8]. While machine learning approaches have been explored, they often lack the precision required for reliable screening [8]. In this context, our research highlights the potential of LLMs to enhance the efficiency and accuracy of literature screening. Our findings suggest that LLMs, with their advanced natural language processing capabilities, can effectively automate significant portions of the screening process, aligning closely with the accuracy of manual methods. This advancement could significantly streamline the preparation of systematic reviews and meta-analyses, offering a promising alternative to traditional methods and addressing the limitations observed with earlier machine learning models.

Our primary analysis revealed that using literature included after conventional manual full-text review as the reference standard, the sensitivity and specificity of LLM-assisted literature screening ranged from 0.77 to 0.99 and 0.91 to 0.98, respectively. Furthermore, the SROC curve, constructed based on the true positive and false positive results from 6 studies, indicated a high level of accuracy for LLM-assisted literature screening, with an AUC ranging from 0.94 to 0.97. Post hoc analysis incorporating modified prompts demonstrated that LLM-assisted literature screening achieved even higher sensitivity (0.98, 95% CI 0.74-1.00) while maintaining a similarly high level of specificity and AUC (0.94-0.99 and 0.99-1.00, respectively). Currently, limited research has explored the accuracy of LLM-assisted screening in meta-analysis for the development of high-level evidence-based medicine. Our research establishes a foundation for the future application of LLMs in the literature screening process of meta-analyses.

Comparison With Previous Work

Our study results demonstrate that LLM-assisted literature screening offers superior accuracy compared with traditional machine-learning models. Previous studies [24,27,29] have reported relatively low sensitivity for machine learning–assisted literature screening, ranging from 0.24 to 0.80. Our study also found that 2 semiautomated tools used for title and abstract screening in literature reviews had relatively low sensitivity, ranging from 0.35 to 0.64. The advantages of LLMs are evident in 3 key aspects. First, large language models possess advanced capabilities in language understanding and text generation, surpassing the capabilities of traditional screening tools [15]. This distinctive feature enables LLMs to excel at identifying relevant literature and discerning irrelevant studies. Conversely, machine learning models require predefined training and validation datasets, including key literature inputs, and often necessitate human review, thereby increasing the barrier to their implementation [24,27]. In contrast, LLMs can rapidly and efficiently generate screening results without the need for training data or human review, owing to their user-friendly conversational interface. This significantly enhances efficiency and reduces workload. Furthermore, research suggests that

LLM-assisted screening can achieve a tenfold reduction in screening time compared with manual screening [1]. While machine learning models also offer time-saving benefits, LLMs eliminate the need for training data and key literature inputs, potentially yielding even greater time savings. Second, previous studies [14,29-31] using machine learning and other natural language processing tools for literature screening reported sensitivities ranging from 0.75 to 0.90, which is consistent with the sensitivity observed in our study. However, our findings indicate that the specificity of LLM-assisted screening (95% CI 0.91-0.98) was notably higher than that reported in previous studies (95% CI 0.69-0.90), suggesting a potential advantage of LLMs in accurately identifying literature relevant to the research topic. While LLMs exhibited very high specificity in both primary, second, and post hoc analyses, it is important to acknowledge that these high-performance estimates might be partially attributed to an overrepresentation of true-negative literature in the datasets used. Third, LLMs exhibit a capacity for continuous learning and self-improvement, akin to human learning processes. With appropriate prompts and instructions, LLMs can refine their performance iteratively, leading to progressively enhanced accuracy. New iterations of LLMs are released approximately every 3-6 months, and these updates are anticipated to further improve sensitivity and specificity during literature screening in meta-analyses. Furthermore, LLMs offer broad applicability and functional extensibility across diverse topics and formats, enabling users to develop customized chatbots tailored to specific research needs. These advantages collectively lower the barrier to entry, reduce workload, and maintain high levels of accuracy, potentially revolutionizing the literature screening process in the future.

In contrast to previous studies [1,9,14] that used a single LLM for literature screening, this study used 3 models concurrently, thereby more accurately reflecting the conventional manual screening process. For instance, Oami et al [1]. relied solely on the ChatGPT-4 Turbo model (released November 7, 2023). Recognizing the ongoing evolution of LLMs, this study expanded the model set to include the updated ChatGPT-4o (released May 13, 2024), Claude-3.5 Sonnet (released June 21, 2024), and Gemini-1.5 Pro (released May 14, 2024). The combined sensitivity and specificity of LLM-assisted literature screening achieved in this study were 0.87 (95% CI 0.77-0.99) and 0.96 (95% CI 0.91-0.98), respectively. These results surpass the sensitivity of 0.75 (95% CI 0.43-0.92) reported by Oami et al [1], suggesting that updated LLMs may enhance screening sensitivity. Specificity remained consistent with the 0.99 (95% CI 0.99-0.99) reported by Oami et al [1], highlighting the potential of LLMs to effectively identify irrelevant literature. While a direct comparison may be limited due to potential heterogeneity introduced by differing research themes, this study provides a novel approach and establishes a foundation for the broader application of LLMs in facilitating literature screening for meta-analyses.

Post hoc analysis revealed that modified prompts significantly improved the sensitivity and specificity of LLM-assisted literature screening to 0.98 (95% CI 0.74-1.00) and 0.98 (95% CI 0.94-0.99), respectively. This underscores the substantial impact of prompt content on LLM performance in literature screening and the quality of meta-analyses. Recent research on prompt engineering has demonstrated the influence of prompts on LLM performance and proposed strategies for tailoring LLM responses to specific topics [17,32,33]. In this study, prompts were designed based on the PICOS framework for each research topic. During the post hoc analysis, false-negative results from studies 1, 2, 3, 5, and 6 were reviewed. This analysis indicated a positive correlation between the complexity of inclusion criteria and the likelihood of LLMs making “exclude” decisions. LLMs exhibited a tendency to strictly adhere to the inclusion criteria specified in the prompts. Conversely, human reviewers typically apply inclusion criteria more conservatively during the initial title and abstract screening phase to minimize the risk of overlooking potentially relevant studies. Based on this observation, it was determined that minor discrepancies between the inclusion criteria in the prompts and the titles or abstracts could be tolerated. Consequently, the prompts were revised to relax the inclusion criteria. In total, 3 iterations of inquiries were then conducted with the modified prompts to optimize sensitivity and reduce false-negative results. The post hoc analysis confirmed an improvement in the sensitivity of LLM-assisted literature screening. However, some false-negative literature was still missed in the final comprehensive analysis, potentially impacting the robustness of the meta-analysis conclusions.

To assess the impact of false-negative literature on the final conclusions, a separate post hoc meta-analysis was conducted. This analysis compared the pooled effect sizes derived from LLM-assisted screening (including only true positives) with those derived from conventional screening (including both true positives and false negatives) for topic studies. The results indicated comparable outcomes between the two methods in most topic studies. Furthermore, the false-negative papers did not substantially alter the overall conclusions of the corresponding topic studies in the majority of instances. Notably, however, for outcomes on lymph node dissection (study 3) and overall survival (study 5), the exclusion of false negatives shifted the results from statistically significant positive effects to nonsignificant negative effects. An examination of the studies included in the study 3 and 5, focusing on the outcomes of lymph node dissection and overall survival, revealed significant heterogeneity and publication bias in the reported findings. This suggests that in meta-analyses demonstrating low heterogeneity and an absence of publication bias, the inclusion or exclusion of potential false-negative studies may not substantially impact the overall conclusions. However, these findings underscore the need for further research on the application of LLMs for literature screening in meta-analyses. Future research should focus on developing prompts that are more readily interpretable by LLMs and exploring techniques for continuous self-correction within LLMs to improve sensitivity.

Limitations

This study acknowledges several limitations. First, its focus is exclusively on meta-analyses within thoracic surgery. Therefore, the generalizability and broader applicability of LLM-assisted literature screening to other fields require further investigation. Future research should evaluate the performance of LLMs across diverse meta-analysis fields to assess their validity for literature screening across various domains of evidence-based medicine.

Second, ongoing updates to LLMs may introduce variations in the quality of model outputs over time, potentially influencing the strength of evidence synthesized in meta-analyses. Third, the use of conventional manual screening as a reference standard may introduce inherent errors in inclusion and exclusion decisions. Existing research [34] indicates that error rates for human reviewers in literature screening range from 6.68% to 21.11% across different fields, with an average error rate of 10.76%. This suggests that the reference standard itself is not infallible, and a comparable error rate for LLM-assisted screening could be considered acceptable. Fourth, a key limitation of our study is the use of previously published systematic reviews, which raises the risk of bias, as the content of these reviews may have been included in the training materials for LLMs. To mitigate this, we redesigned the search strategy and replicated the inclusion criteria and timeframes outlined in these meta-analyses for a new round of literature retrieval and screening. In addition, we evaluated the results with new readers to ensure the integrity of our findings. Fifth, an additional key limitation of our study lies in the method of accessing LLMs, as we used a web-based interface instead of an application programming interface (API). While convenient, web-based access lacks the flexibility, performance, offline functionality, and data security provided by API-based deployment. This reliance on external servers outside the researchers' control may have introduced some bias in the literature screening process. Future studies should explore API-based access, which enables local or server deployment, offering greater control, enhanced security, and better integration with research workflows. APIs also allow secure handling of sensitive data and more efficient operation in offline or resource-constrained settings. Although the web-based interface was sufficient for this study, adopting API-based access in future research could address these limitations and improve reliability and security.

Despite these limitations, the integration of LLMs into the meta-analysis workflow represents a significant advancement with the potential to enhance productivity and accelerate the speed and quality of resource and knowledge synthesis.

Conclusions

LLM-assisted screening, particularly at the full-text screening level and with revised prompts, can achieve accuracy comparable with manual screening. This suggests LLMs hold significant potential for streamlining literature screening in systematic reviews, reducing workload without sacrificing quality. Integrating LLMs into evidence synthesis workflows could accelerate the production of high-quality reviews, facilitating more timely translation of research into practice and policy.

Acknowledgments

We thank Lei Peng, Xing-Yu Liu, and Xu-Yang Wang for their assistance in conducting conventional manual literature screening and for providing signed statements agreeing to be cited in our paper. This study was supported by the Key Research Project of Sichuan Province (2023YFS0199), the Chengdu Science and Technology Project (2022-YF05-01833-SN), and the Sichuan Cadre Health Care Research Project (GBKT23020). The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

All data generated or analyzed during this study are included in this published article and Multimedia Appendices 1 and 2.

Authors' Contributions

Z-YD contributed to conceptualization, methodology, software, formal analysis, writing, and original draft preparation. CS contributed to data curation, writing, and original draft preparation. F-QW contributed to writing, reviewing, editing, and project administration. Y-LJ contributed to formal analysis, visualization, and investigation. Z-YL contributed to data curation, writing, and original draft preparation. YW contributed to supervision, software, and validation. QP contributed to writing, reviewing, editing, and supervision. All authors read and approved the final manuscript. Z-YD, CS, and F-QW contributed equally to this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional tables and figures.
[DOCX File , 5484 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Additional material content.
[DOCX File , 24 KB-Multimedia Appendix 2]

Multimedia Appendix 3

CONSORT-EHEALTH Checklist.
[PDF File (Adobe PDF File), 22394 KB-Multimedia Appendix 3]

References

1. Oami T, Okada Y, Nakada T. Performance of a large language model in screening citations. JAMA Netw Open. 2024;7(7):e2420496. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.20496] [Medline: 38976267]
2. Harris JD, Quatman CE, Manring M, Siston RA, Flanigan DC. How to write a systematic review. Am J Sports Med. 2014;42(11):2761-2768. [doi: 10.1177/0363546513497567] [Medline: 23925575]
3. Courtney DB, Bennett K, Szatmari P. The forest and the trees: evidence-based medicine in the age of information. J Am Acad Child Adolesc Psychiatry. 2019;58(1):8-15. [doi: 10.1016/j.jaac.2018.06.035] [Medline: 30577942]
4. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700. [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
5. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017;7(2):e012545. [FREE Full text] [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
6. Peery AF, Shaukat A, Strate LL. AGA clinical practice update on medical management of colonic diverticulitis: expert review. Gastroenterology. 2021;160(3):906-911.e1. [FREE Full text] [doi: 10.1053/j.gastro.2020.09.059] [Medline: 33279517]
7. Le Glaz A, Haralambous Y, Kim-Dufor D, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res. 2021;23(5):e15708. [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]
8. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. 2020;46(3):383-400. [FREE Full text] [doi: 10.1007/s00134-019-05872-y] [Medline: 31965266]
9. Blaizot A, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins M, et al. Using artificial intelligence methods for systematic review in health sciences: a systematic review. Res Synth Methods. 2022;13(3):353-362. [doi: 10.1002/jrsm.1553] [Medline: 35174972]
10. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med. 2023;6(1):210. [FREE Full text] [doi: 10.1038/s41746-023-00958-w] [Medline: 37973919]
11. Shibue K. Artificial intelligence and machine learning in clinical medicine. N Engl J Med. 2023;388(25):2398. [doi: 10.1056/NEJMc2305287] [Medline: 37342936]
12. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med. 2023;29(10):2396-2398. [doi: 10.1038/s41591-023-02412-6] [Medline: 37391665]
13. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large language model capabilities in perioperative risk prediction and prognostication. JAMA Surg. 2024;159(8):928-937. [doi: 10.1001/jamasurg.2024.1621] [Medline: 38837145]
14. Kohandel Gargari O, Mahmoudi MH, Hajisafarali M, Samiee R. Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. BMJ Evid Based Med. 2024;29(1):69-70. [FREE Full text] [doi: 10.1136/bmjebm-2023-112678] [Medline: 37989538]
15. 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. 2023;12(1):72. [FREE Full text] [doi: 10.1186/s13643-023-02243-z] [Medline: 37120563]
16. Winckelmans T, Decaluwé H, De Leyn P, Van Raemdonck D. Segmentectomy or lobectomy for early-stage non-small-cell lung cancer: a systematic review and meta-analysis. Eur J Cardiothorac Surg. 2020;57(6):1051-1060. [doi: 10.1093/ejcts/ezz339] [Medline: 31898738]
17. Zeng W, Zhang W, Zhang J, You G, Mao Y, Xu J, et al. Systematic review and meta-analysis of video-assisted thoracoscopic surgery segmentectomy versus lobectomy for stage I non-small cell lung cancer. World J Surg Oncol. 2020;18(1):44. [FREE Full text] [doi: 10.1186/s12957-020-01814-x] [Medline: 32106856]
18. Zhang J, Feng Q, Huang Y, Ouyang L, Luo F. Updated evaluation of robotic- and video-assisted thoracoscopic lobectomy or segmentectomy for lung cancer: a systematic review and meta-analysis. Front Oncol. 2022;12:853530. [FREE Full text] [doi: 10.3389/fonc.2022.853530] [Medline: 35494020]
19. Righi I, Maiorca S, Diotti C, Bonitta G, Mendogni P, Tosi D, et al. Oncological outcomes of segmentectomy versus lobectomy in clinical stage I non-small cell lung cancer up to two centimeters: systematic review and meta-analysis. Life (Basel). 2023;13(4):947. [FREE Full text] [doi: 10.3390/life13040947] [Medline: 37109476]

20. Zhang W, Chen S, Lin X, Chen H, He R. Lobectomy versus segmentectomy for stage IA3 (T1cN0M0) non-small cell lung cancer: a meta-analysis and systematic review. Front Oncol. 2023;13:1270030. [FREE Full text] [doi: 10.3389/fonc.2023.1270030] [Medline: 37849809]
21. Lin H, Peng Z, Zhou K, Liang L, Cao J, Huang Z, et al. Differential efficacy of segmentectomy and wedge resection in sublobar resection compared to lobectomy for solid-dominant stage IA lung cancer: a systematic review and meta-analysis. Int J Surg. 2024;110(2):1159-1171. [FREE Full text] [doi: 10.1097/JS9.0000000000000896] [Medline: 37983767]
22. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210. [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
23. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. 2023;25:e50638. [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
24. Oami T, Okada Y, Sakuraya M, Fukuda T, Shime N, Nakada T. Efficiency and workload reduction of semi-automated citation screening software for creating clinical practice guidelines: a prospective observational study. J Epidemiol. 2024;34(8):380-386. [FREE Full text] [doi: 10.2188/jea.JE20230227] [Medline: 38105001]
25. ASReview. URL: https://asreview.nl/download/ [accessed 2024-12-10]
26. Abstrackr. URL: http://abstrackr.cebm.brown.edu/ [accessed 2024-12-12]
27. Gates A, Guitard S, Pillay J, Elliott SA, Dyson MP, Newton AS, et al. Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. Syst Rev. 2019;8(1):278. [FREE Full text] [doi: 10.1186/s13643-019-1222-2] [Medline: 31727150]
28. Ting YT, Hsieh TC, Wang YF, Kuo YC, Chen YJ, Chan PK, et al. Performance of ChatGPT incorporated chain-of-thought method in bilingual nuclear medicine physician board examinations. Digit Health. 2024;10:20552076231224074. [FREE Full text] [doi: 10.1177/20552076231224074] [Medline: 38188855]
29. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021;3(2):125-133. [doi: 10.1038/s42256-020-00287-7]
30. Perlman-Arrow S, Loo N, Bobrovitz N, Yan T, Arora RK. A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Res Synth Methods. 2023;14(4):608-621. [doi: 10.1002/jrsm.1636] [Medline: 37230483]
31. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the abstrackr machine learning tool. Syst Rev. 2018;7(1):45. [FREE Full text] [doi: 10.1186/s13643-018-0707-8] [Medline: 29530097]
32. Gruda D. Three ways ChatGPT helps me in my academic writing. Nature. 2024. [doi: 10.1038/d41586-024-01042-3] [Medline: 38589655]
33. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. Ann Biomed Eng. 2023;51(12):2629-2633. [doi: 10.1007/s10439-023-03272-4] [Medline: 37284994]
34. Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. PLoS One. 2020;15(1):e0227742. [FREE Full text] [doi: 10.1371/journal.pone.0227742] [Medline: 31935267]

Abbreviations

API: application programming interface
AUC: area under the curve
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
LLM: large language model
PICOS: Population, Intervention, Control, Outcome, and Study
SROC: summary receiver operating characteristic curve
STARD: Standards for Reporting of Diagnostic Accuracy Studies

Edited by T de Azevedo Cardoso; submitted 13.10.24; peer-reviewed by J Valverde-Rebaza, CH Krag; comments to author 02.12.24; revised version received 14.12.24; accepted 20.02.25; published 11.03.25

Please cite as:
Dai Z-Y, Wang F-Q, Shen C, Ji Y-L, Li Z-Y, Wang Y, Pu Q
Accuracy of Large Language Models for Literature Screening in Thoracic Surgery: Diagnostic Study
J Med Internet Res 2025;27:e67488
URL: https://www.jmir.org/2025/1/e67488
doi: 10.2196/67488
PMID: