BMC Medical Research Methodology

RESEARCH

Open Access

Streamlining systematic reviews with large language models using prompt engineering and retrieval augmented generation

Fouad Trad[1*], Ryan Yammine[2], Jana Charafeddine[1], Marlene Chakhtoura[2], Maya Rahme[2], Ghada El-Hajj Fuleihan[2] and Ali Chehab[1]

Abstract

Background Systematic reviews (SRs) are essential to formulate evidence-based guidelines but require time-consuming and costly literature screening. Large Language Models (LLMs) can be a powerful tool to expedite SRs.

Methods We conducted a comparative study to evaluate the performance of a commercial tool, Rayyan, and an in-house LLM-based system in automating the screening of a completed SR on Vitamin D and falls. The SR retrieved 14,439 articles, and Rayyan was trained with 2,000 manually screened articles to categorize the rest as most likely to exclude/include, likely to exclude/include and undecided. We analyzed Rayyan's title/abstract screening performance using different inclusion thresholds. For the LLM, we used prompt engineering for title/abstract screening and Retrieval-Augmented Generation (RAG) for full-text screening. We evaluated performance using article exclusion rate (AER), false negative rate (FNR), specificity, positive predictive value (PPV), and negative predictive value (NPV). Additionally, we compared the time required to complete screening steps of the SR using both approaches against the manual screening method.

Results Using Rayyan, including considered as undecided or likely to include for title/abstract screening resulted in an AER of 72.1% and an FNR of 5%. The total estimated screening time, including manual review of articles flagged by Rayyan, was 54.7 hours. Lowering the Rayyan threshold to 'likely to exclude' reduced the FNR to 0% and the AER to 50.7%, but increased the screening time to 81.3 h. Using the LLM system, after title/abstract and full-text screening, 78 articles remained for manual review, including all 20 identified by traditional methods. The LLM achieved an AER of 99.5%, specificity of 99.6%, PPV of 25.6%, and NPV of 100%, with a total screening time of 25.5 h, including manual review of the 78 articles, reducing the manual screening time by 95.5%.

Conclusions The LLM-based system significantly enhances SR efficiency, compared to manual methods and Rayyan while maintaining low FNR.

Keywords Systematic reviews, Large language models, Prompt engineering, Retrieval-augmented generation, Rayyan AI

*Correspondence:
Fouad Trad
fat10@mail.aub.edu
[1]Department of Electrical and Computer Engineering, American University of Beirut, Beirut, Lebanon
[2]Faculty of Medicine, American University of Beirut, Beirut, Lebanon

Background

Systematic Reviews (SRs) are an essential pillar for evidence-based guideline development. However, their process is labor-intensive and time-consuming, requiring authors to screen thousands of articles, with SRs taking on average 67.3 weeks to complete [1]. One of the most time-consuming steps in SRs is literature screening, which is conducted in duplicate and independently, in two steps: title/abstract screening followed by full-text screening.

There is an increasing demand for rapid and frequent SRs by scientists to stay up-to-date in their field as scientific data output is increasing rapidly worldwide, with the corpus of literature doubling every 9 years [2]. Other pressing needs are incurred by pandemics and living practice guidelines.

To meet this increasing demand, many artificial intelligence-based tools have emerged in an attempt to expedite this process, such as Abstrackr, Rayyan AI, ASreviews, Colandr, and DistillerAI [3–7]. These tools vary significantly in their core algorithm and features, but are mostly limited to title/abstract screening. In one review conducted in 2020 comparing various AI-based title/abstract screening tools, Rayyan AI scored the highest in weighted feature analysis [8]. Rayyan AI is a web-based semi-automated screening tool, developed by Qatar Computing Research Institute [4]. It works by feeding words, pairs of words and Medical Subject Headings (MeSH) terms from the titles and abstracts to a Machine Learning (ML) algorithm, more specifically, a Support Vector Machine (SVM) classifier [4].

Recent studies have shown success in the use of Large Language models (LLMs) such as GPT-3.5 and GPT-4 in title/abstract screening [9, 10]. Attempts to leverage LLMs for both title/abstract and full text screening are limited [11]. We investigate the use of LLM techniques, such as Prompt Engineering and Retrieval-Augmented Generation (RAG), to automate the aforementioned processes. We propose an end-to-end system powered by GPT-4 that receives an article along with inclusion and exclusion criteria, and then decides whether to include or exclude the article from the SR.

We capitalize on a completed SR on vitamin D and falls to compare the performance of the two ML-based systems, with the traditional manual method as the gold standard [12].

Methods

Data Preparation

We used data from a recently completed umbrella review on Vitamin D and Falls [12]. After title/abstract screening, 1,680 full-text papers were reviewed, with 20 SRs of Randomized Controlled Trials (RCTs) included in the final review (Appendix 1).

Manual traditional title/abstract screening followed a validated screening guide (Appendix 2). However, results for 430 articles were inadvertently not saved, reducing the total dataset to 17,346 articles. Importantly, none of the 430 excluded articles were among the final 20 articles included in the completed analysis of the completed SR using the gold standard method [12].

The 17,346 articles were imported into Rayyan software. Duplicates were removed using Rayyan's Duplicate Detection Tool.

Rayyan AI

One reviewer trained Rayyan AI by manually screening 2,000 random articles in batches of 100, using the completed umbrella review's title/abstract screening guide (Appendix 2). The reviewer assigned one reason for exclusion for manually excluded articles, following the screening guide.

After each set of 100 articles screened, Rayyan would classify unscreened articles into five categories: "Most Likely To Exclude", "Likely To Exclude", "Undecided", "Likely To Include" or "Most Likely To Include" based on patterns learned during the screening phase. We considered articles rated as "Undecided" or higher to require further manual title/abstract screening, and excluded articles rated as "Likely To Exclude" or lower (Threshold A). Furthermore, we analyzed the results of lowering the threshold for exclusion to "Most Likely To Exclude" only (Threshold B).

A second researcher intentionally excluded the 20 final articles identified with the gold standard method [12] from the batch to be screened for Rayyan model training. This was to ensure that he/she did not accidentally include any of these articles in the model training, which could subsequently affect the integrity of the comparison between the manual and the Rayyan methods. This approach does not apply when deploying Rayyan to de novo systematic reviews.

After training the model, we evaluated the performance of Rayyan on all unscreened articles including the above 20 articles selected in the completed manual SR [12]. We stopped training Rayyan when the number of unscreened articles in each category stabilized - meaning that after each batch of 100 screened articles, the distribution of articles across the five categories showed minimal change. For this study, this saturation occurred at 2000 articles. (Fig. 1).

Rayyan AI has recently introduced a full-text screening feature. However, this feature does not incorporate AI, automation, or machine learning. Instead, it functions as a platform for users to manually review and record their screening decisions without assisting or learning from the process. For this reason, we will limit our evaluation of Rayyan AI to title and abstract only.

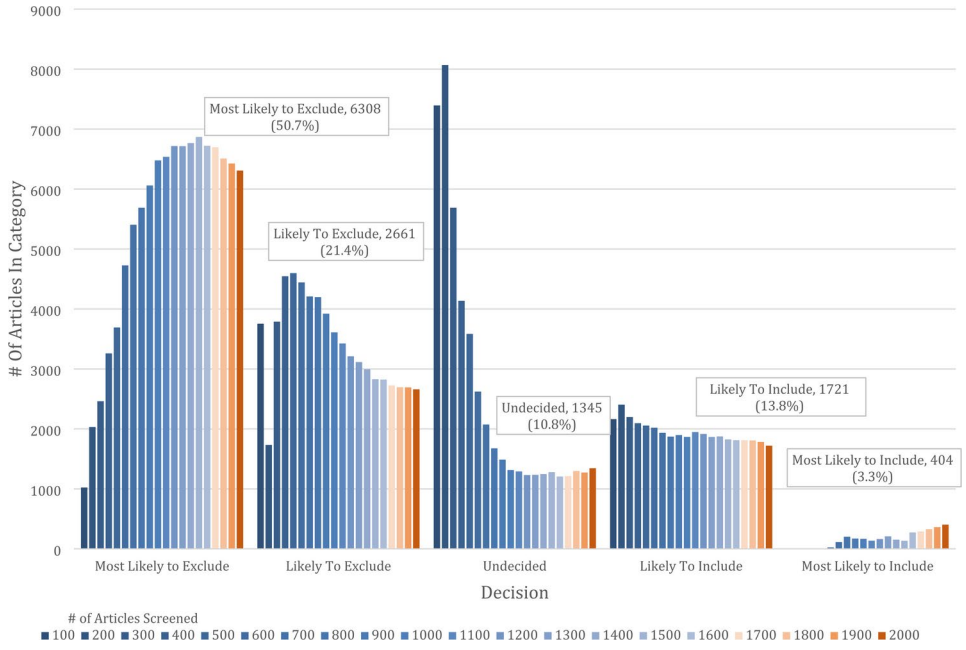

Fig. 1 Rayyan classification of unscreened articles into its five default categories with increasing training*. *Training is done in batches of 100, reaching a total of to 2,000 Articles

Proposed LLM-based approach

The proposed system relies on two LLM techniques: prompt engineering, which involves designing specific input prompts to guide the model's responses, and RAG, which combines external data retrieval with generative capabilities to enhance accuracy and relevance. Together, these techniques automate the two screening phases: Title and Abstract Screening, and Full-Text Screening. We called the GPT-4 model via the OpenAI API (using 'gpt-4' as the model's name), providing structured prompts and receiving responses programmatically. The prompts were designed to ensure the model responded in a specific format, maintaining consistency and minimizing variability in decision-making.

- In title/abstract screening (Phase 1), we input the titles and abstracts of articles into GPT-4 for screening. We give the model a system prompt that instructs it to act as a professional medical researcher performing title/abstract screening. This system prompt helps the model adopt the specified role and respond with the appropriate level of expertise and focus, improving accuracy and consistency during screening (Prompt Engineering). Then, we prompt the model with a series of questions identical to the traditional screening criteria used in the original article's title/abstract screening guide (Appendix 2). The model responds to each question with "yes," "no," or "unsure." When the model is certain about its decision on an article, we proceed accordingly. If the model is uncertain, we retain the article, just as we do with the traditional process, improving sensitivity.
- Articles that pass the first phase undergo a more thorough full-text screening (Phase 2), employing RAG. The full-text PDFs were first obtained manually and then processed using a Python script that stores them in a vector store using LlamaIndex for efficient retrieval during screening. Here, the full text of each article serves as the document set from which the GPT-4 model retrieves information. A new set of questions identical to the ones used for traditional screening (Appendix 3) is used to evaluate the full texts. The model's responses in this phase to the first five questions are categorized as "yes," "no," or "unsure." Articles are included or excluded similarly to Step 1. The final question prompts the model to identify the outcome studied in the review—falls, fractures, or mortality. The article is only included if "Falls" is one of the outcomes.

The prompts used for both phases can be found in Appendix 4.

To enhance transparency and facilitate an effective review process, the outcomes of all questions are automatically documented in an Excel sheet during both phases for every article (Appendix 5). Since the prompts asked the model to respond in a structured format, a Python script was used to log the model's answers directly into the sheet, eliminating any manual intervention in transferring results. This ensures a fully automated workflow, removing the possibility of any sort of

error. This logging method enables reviewers to assess the rationale behind the model's decisions. This structured documentation ensures that all decisions are traceable and reviewable. This provides a clear audit trail and supports any necessary re-evaluation of articles automatically screened by the model.

Statistical analysis

We considered the completed SR on Vitamin D and Falls as our gold standard for comparison. For both steps, true positives were defined as articles correctly included for further screening, as they were among the articles included for final analysis, and true negatives were as articles correctly excluded. False positives were articles included by the model for further manual screening but excluded not ultimately included after traditional full-text screening using the manual method, while false negatives were articles excluded by the model but included after manual traditional full-text screening. We used these defined values to calculate the performance metrics described below.

For title/abstract screening using both methods, we evaluated false negative rate (FNR) and article exclusion rate (AER). AER is defined as the total number of articles automatically excluded during a step divided by the total number of articles at the beginning of the relevant step, as illustrated in the equation below:

$$AER_{\text{Title/Abstract}}(\%) = \frac{\text{Number of automatically excluded articles during Title/Abstract}}{\text{Total number of articles at the beginning of Title/Abstract}} \times 100$$

For full-text screening, which was assessed using the LLM model only (since Rayyan's semi-automation tool does not support this step), we evaluated FNR, AER, specificity, positive predictive value (PPV), and negative predictive value (NPV). We also assessed these performance metrics from start to end (title/abstract, and full text screen) using the LLM approach. For full-text screening, AER was calculated as illustrated below:

$$AER_{\text{Full-Text}}(\%) = \frac{\text{Number of automatically excluded articles during Full-Text}}{\text{Total number of articles at the beginning of Full-Text}} \times 100$$

To estimate workload reduction, we considered both the AER and the time taken to complete screening of the remaining articles. Additionally, FNR was calculated to assess the risk of erroneously excluding relevant articles.

We estimated time required for each screening method as follows:

For the traditional screening method, we estimated the time required for both title/abstract screening (M1) and full-text screening (M2).

For title/abstract screening using Rayyan AI, we calculated the time taken to train the model (R1) and the time needed for manual title/abstract screening of articles remaining after automatic screening (R2). The total time for this process was R = R1 + R2. We estimated M1, R1, and R2 based on the time it took the reviewer to screen 100 articles for Rayyan's training.

For the LLM-based model, we recorded time for automatic title/abstract screening (S1) and full-text screening (S2). Additionally, we estimated time required for manual full-text screening of articles remaining after the automatic process (SM). SM and M2 were calculated based on our team's experience, which estimated that manually screening one full-text article takes an average of 15 min.

The total time required to complete title/abstract, and full-text screening using the LLM system was S1 + S2 + SM, where SM is equal to 15 min multiplied by the number of remaining articles. Since our primary focus was on screening time rather than document retrieval, we did not include the time required to collect and prepare full-text PDFs in any method. This ensures a fair comparison, as retrieval would be a necessary preliminary step regardless of the approach used.

Results

Rayyan title/abstract screening

Of the original 17,346 articles, 2,907 articles were deleted after duplicate removal, and 14,439 remained. The reviewer took approximately 1 h to perform title/abstract screening on 100 articles. Screening all 14,439 articles would take them approximately M1 = 144.4 h.

Of the 2,000 articles screened manually to train Rayyan, 1,727 (86.35%) were excluded, and 273 (13.65%) were included. This step took approximately R1 = 20 h. Of the remaining 12,439 unscreened articles, Rayyan classified 6,308 (50.7%) as most likely to exclude, 2,661 (21.4%) likely to exclude, 1,345 (10.8%) undecided, 1,721 (13.8%) likely to include and 404 (3.3%) most likely to include (Fig. 1). Of the 20 articles included for final analysis in our traditional manual method: 3 were ranked as most likely to include, 6 were ranked as likely to include, 10 were ranked as undecided and 1 was ranked as likely to exclude.

When using Threshold A, 8,969 (72.1%) of the 12,439 unscreened articles were excluded, with the remaining 3,470 (27.9%) articles to be screened manually (Fig. 2A), with 1 false negative result. This resulted in: AER 72.1%, FNR 5%, reducing the time needed to complete title/abstract screening using Rayyan to 54.7 h, an 89.7-hour reduction (62%) when compared to the traditional methods (Fig. 2A).

In contrast, when using Threshold B, excluded articles decreased to 6,308 (50.7%), and the remaining 6,131 (49.3%) would undergo further manual screening (Fig. 2B), with no false negative results. This inclusion threshold resulted in an AER of 50.7%, and an FNR of 0%, reducing the time needed to complete title/abstract screening using Rayyan to 81.3 h, a 63.1-hour reduction

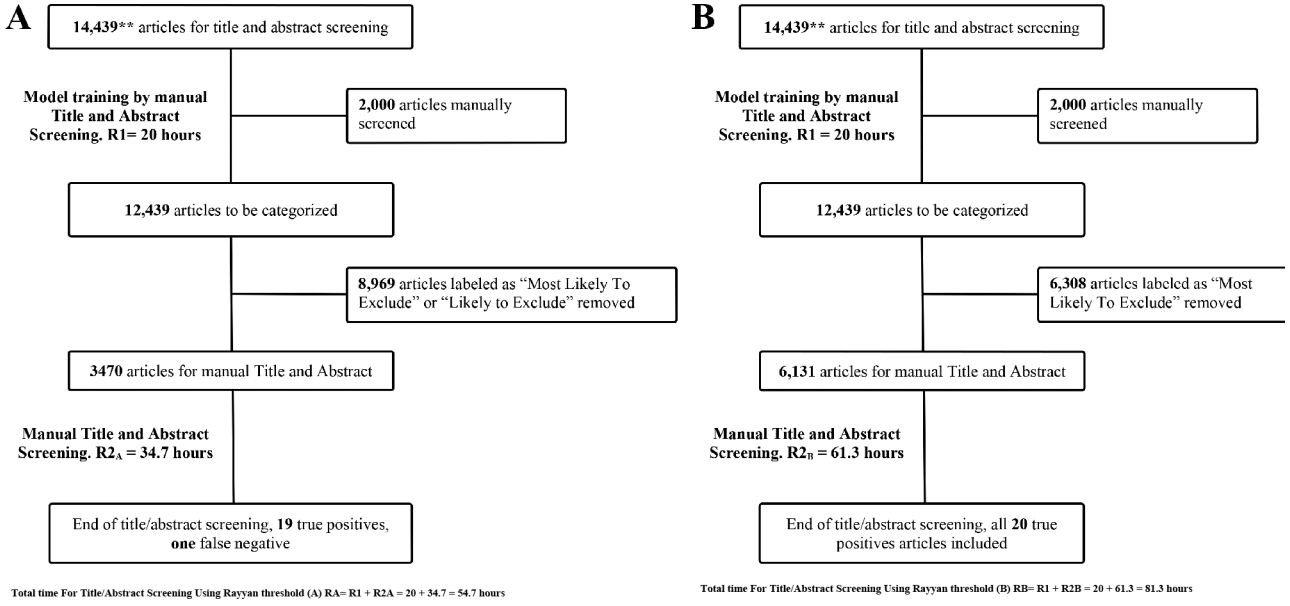

Fig. 2 **A**: Flow diagram of Rayyan title/abstract screening steps and results using Threshold A* as inclusion criteria. *"Undecided" as threshold for inclusion. **Of the original 17,776 citations, 430 articles were excluded as their results were inadvertently not saved. 2,907 articles were deleted after duplicate removal, of the remaining 17,346 articles and 14,439 remained. **B**: Flow diagram of Rayyan title/abstract screening steps and results using Threshold B* as inclusion criteria. *"Likely To Exclude" as threshold for inclusion. **Of the original 17,776 citations, 430 articles were excluded as their results were inadvertently not saved. 2,907 articles were deleted after duplicate removal, of the remaining 17,346 articles and 14,439 remained

(44%) when compared to the traditional methods (Fig. 2B).

LLM title/abstract, and full-text screening

Of the 14,439 articles processed by the GPT-4 model for title/abstract, 3,298 articles (22.8%) met the inclusion criteria and advanced to Phase 2, achieving an AER of 77.2%. None of the 20 retained in the traditional method were excluded, achieving an FNR of 0%. This step took S1 = 2 h to run.

In the subsequent RAG-based full-text screening phase, the 3,298 full-text articles were evaluated. Out of these, only 78 articles (or 2.37%) were included for manual review, including all 20 articles retained in the traditional method. This step required S2 = 4 h to run, compared to M2 = 1680*15 minutes = 420 h for the traditional method. The metrics for this step are as follows: AER: 97.63%, specificity 99.6%, PPV 25.6%, and NPV 100%.

For the entire process, including both phases, the LLM method achieved the following metrics: AER 99.5%, specificity 99.6%, PPV 25.6%, and NPV 100%. Manual screening of the remaining 78 articles would take approximately 19.5 h (SM), bringing the total time for title/abstract, and full-text screening using the LLM approach to 25.5 h (Fig. 3). This represents a time reduction of 538.9 h (95.5%) compared to the traditional method, which required an estimated M1 + M2 = 564.4 h.

A summary of the performance for both approaches can be found in Table 1.

Discussion

Our study shows that both Rayyan AI and the LLM-based system dramatically reduced the workload for SRs compared to traditional methods, while maintaining a low FNR. However, the LLM-based system stood out by not only automating title/abstract screening but also incorporating full-text screening, a more challenging task, through advanced techniques like prompt engineering and RAG. This enabled the LLM to reduce the number of articles for manual full-text review to just 78 out of the original 14,439.

Crucially, the LLM-based system achieved a 95.5% reduction in screening time compared to the traditional method, from 564.4 h using the traditional approach, to only 25.5 h. Even more importantly, the LLM maintained a perfect FNR of 0%, meaning no relevant articles were missed during screening. Unlike Rayyan and traditional methods, which rely on human input, the LLM system drastically reduces human intervention, lowering the risks of human error and bias. This impressive combination of time savings and accuracy highlights the LLM's transformative potential for making SRs more efficient and reliable.

According to the Cochrane Collaboration, literature screening ideally involves two reviewers who independently screen articles by following strict screening

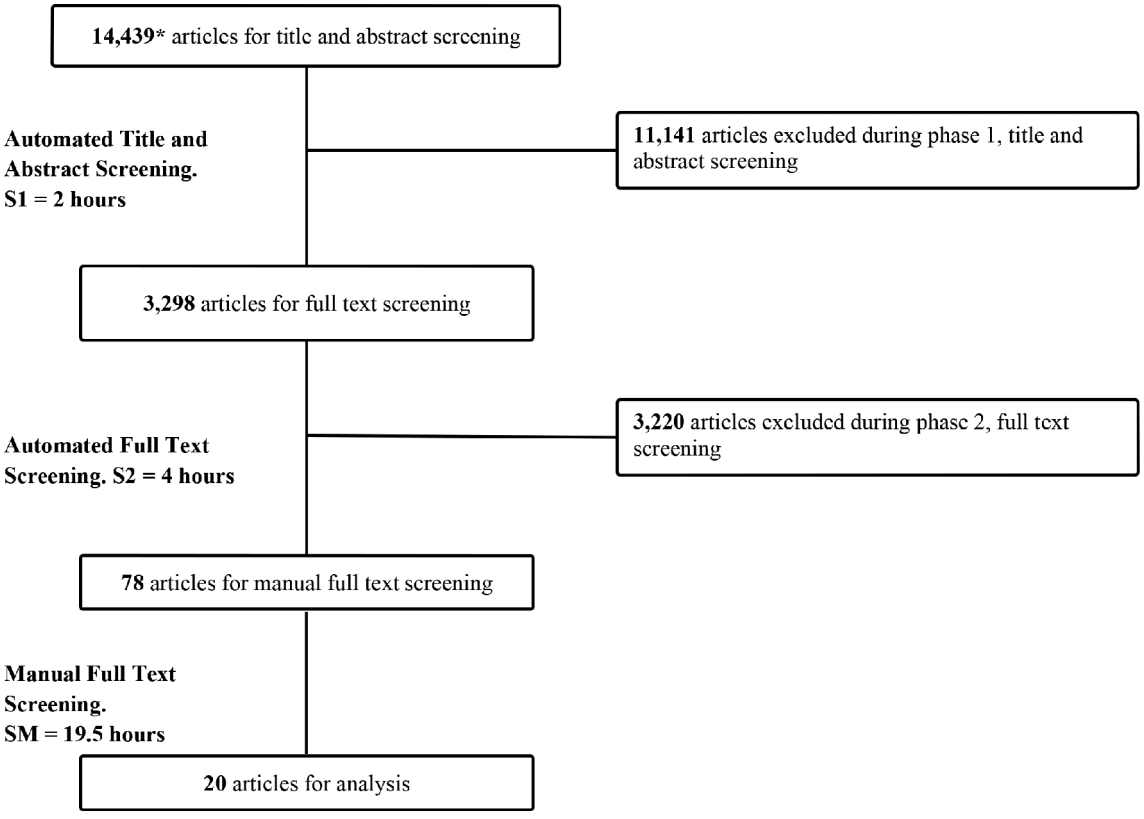

Fig. 3 Flow diagram of LLM Title/abstract and full text screening steps and results. *Of the original 17,776 citations, 430 articles were excluded as their results were inadvertently not saved. 2,907 articles were deleted after duplicate removal, of the remaining 17,346 articles and 14,439 remained

criteria, to minimize bias, maximize sensitivity, ensuring that no important articles are missed [13]. However, this process is highly time-consuming, particularly for large-scale systematic reviews that involve thousands of articles. Given this challenge, automated assistance in the screening process can be beneficial. The LLM-based approach can serve as an initial screening tool, significantly reducing the number of articles that require manual review. Since our results indicate that the LLM system achieves a false negative rate (FNR) of 0, it ensures that no relevant articles are erroneously excluded at this stage. However, this does not mean that human reviewers should be replaced. Instead, after the LLM performs the initial screening, reviewers can focus their efforts on evaluating the remaining articles in the same way as traditional screening methods. By reducing the initial workload, this approach allows researchers to dedicate more time to the final selection process, ultimately streamlining systematic review workflows while maintaining high sensitivity and accuracy.

While few publications have explored the potential of Rayyan software in expediting title/abstract screening, they suffered several drawbacks [14–16]. These include using smaller datasets, with samples varying between 500 and 1512 articles, and lacking details on thresholds used, rendering an assessment of their performance metrics challenging [16]. Nevertheless, our results based on a larger dataset show a similar high sensitivity using Threshold A [14, 15]. However, unlike Valizadeh et al., our study also analyzed a more conservative threshold where false negatives were eliminated [14].

Beyond commercial systems, such as Rayyan, there has been a growing interest in leveraging LLMs to enhance various stages of SRs [17]. Reason et al. evaluated the potential of LLMs to automate tasks such as data extraction, script creation, and report generation within SRs [18]. Others have explored the potential of LLMs in assessment of the quality and risk-of-bias of publications, with varying degrees of success [19–21].

Few publications explored the use GPT-4 to automate title/abstract screening, similar to Phase 1 of our LLM model [10, 22–25]. While these studies demonstrated acceptable performance and time savings, they did not extend to full-text screening—a critical and time-consuming phase of SRs. To our knowledge, the only exception is the work of Khraisha et al. [11]. However,

Table 1 Summary and comparison of the manual method, Rayyan thresholds A and B and the LLM method

		Manual	Rayyan Threshold A	Rayyan Threshold B	LLM
Title/Abstract	Articles to screen*	14,439	14,439	14,439	14,439
	Inclusion Threshold	-	"Undecided"	"Likely to Exclude"	-
	Articles Remaining after automated screening (AER)	-	3,470 (72.1%)	6,131 (50.7%)	3,280 (77.2%)
	Total Articles to Manually Screen	14,439	5,470	8,131	
	Time taken for all Manual Screening Articles	144.4 h	54.7 h	81.3 h	-
	Time for automated screening	-	-	-	2 h
	True Positives (FNR)	N/A (Gold Standard)	19 (5%)	20 (0%)	20 (0%)
	Total time for Step	144.4 h	54.7 h	81.3 h	2 h
	Total Time Saved compared to manual method (%)	N/A (Gold Standard)	89.7 h (62.1%)	63.1 h (43.7%)	-
Full Text**	Articles to screen	1,680	-	-	3,280
	Time to run automated screening	-	-	-	4 h
	Articles Remaining (AER)	-	-	-	78 (97.6%)
	Time to manually screen remaining articles	420 h	-	-	19.5 h
	True Positives (FNR)	N/A (Gold Standard)	-	-	20 (0%)
	Total time for step (hours)	420 h	-	-	23.5 h
Total	Total Time for both steps	564.4 h	-	-	25.5 h
	Total Time Saved compared to manual method (%)	N/A (Gold Standard)	-	-	538.9 h (95.5%)

AER: Article Exclusion Rate, FNR: False Negative Rate

*Of the original 17,776 citations, 430 articles were excluded as their results were inadvertently not saved. 2,907 articles were deleted after duplicate removal, of the remaining 17,346 articles and 14,439 remained

**Rayyan was excluded from full-text comparison, as its article classification feature is not yet supported in its full text screening platform

their method relied on article segmentation for full-text screening. This approach can affect model performance, as it may struggle to grasp the context when processing segmented parts in isolation, contrary to our RAG framework [11]. As a result, it achieved a low sensitivity of 0.42 and 0.38 during phases 1 and 2, respectively [11]. Importantly, these metrics were based on a limited number of citations screened, 300 titles/abstracts and 150 full texts [11]. In contrast, our system handled a larger dataset of 14,439 articles in the title/abstract screening phase, achieving an FNR of 0% (sensitivity of 100%) during both steps, with a high AER. While our study involved a large dataset, the performance of the LLM was not influenced by the number of articles screened, as we did not train or fine-tune the model. Instead, factors such as prompt clarity and retrieval effectiveness in the RAG phase played a larger role.

None of the discussed publications, including ours, assessed the time needed for the development and finalization of the title/abstract and full text screening sheets. This is an iterative and necessary process with a calibration phase implemented before the sheets are ready for use by any of the three methods. This, however, does not affect our comparisons between methods. Our study has several strengths. It implemented testing over 14,000 articles to pilot our approach, as opposed to a maximum of 5,634 in other studies also using LLMs [10, 22–25]. Additionally, it demonstrated strong performance, with an AER of 99.5%, specificity of 99.6%, PPV of 25.6%, and NPV of 100%, outperforming comparable studies in the literature. Although the LLM-based system requires engineering expertise to build the model, once operational, users can easily interact with it by inputting their inclusion and exclusion criteria in the form of questions. This usability feature underscores the practical application of the system in streamlining the review process. Additionally, the transparent logging of each question's outcome in an Excel sheet not only enhances the system's integrity but also facilitates manual subsequent checks of any article, allowing users to trace decisions back to specific responses, thus reinforcing trust in this approach.

While our results demonstrate strong performance, LLMs are not without limitations. One key disadvantage is their dependency on prompt design—suboptimal prompts can lead to inconsistencies in responses. Additionally, LLMs may struggle with complex or nuanced inclusion/exclusion criteria that require deep domain expertise, necessitating careful human oversight.

Although the LLM approach demonstrated significant improvements compared to traditional methods and Rayyan, its performance should be validated across diverse and complex systematic reviews to confirm its robustness and generalizability. Previous research has shown that LLM performance can vary depending on the topic and dataset used [26]. This variation suggests that while our approach achieved strong results in this study, further evaluations across different domains are necessary to ensure consistent performance. Additionally,

future enhancements should focus on refining the logging features to provide even more detailed explanations for each question (knowing why the response was yes, no, or unsure), enhancing explainability and the ability to audit this approach.

Conclusions

Our study demonstrates that the proposed LLM-based system significantly enhances the efficiency of the SR process compared to both traditional methods and the commercially available Rayyan system, while maintaining low FNR. Its excellent performance metrics, ease of use, explainability, alignment with traditional methods, and its time efficiency, position it as a very promising approach. Future work could explore expanding the system's capabilities to support more complex review steps, such as data extraction and synthesis.

Abbreviations

SR	Systematic Review
LLM	Large Language Model
RAG	Retrieval-Augmented Generation
PPV	Positive predictive value
NPV	Negative predictive value NPV
AER	Article Exclusion Rate
FNR	False Negative Rate
RoB	Risk of Bias
MeSH	Medical Subject Headings
SVM	Support Vector machine
RCT	Randomized Controlled Trial
ML	Machine Learning

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-025-02583-5.

Supplementary Material 1

Acknowledgements

Ryan Yammine would like to acknowledge the training received under the Scholars in HeAlth Research Program (SHARP) that was in part supported by the Fogarty International Center and Office of Dietary Supplements of the National Institutes of Health (Award Number D43 TW009118). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

FT, GEHF and AC contributed to the Conceptualization. FT, JC, MC, MR and GEHF contributed to the Formal Analysis. FT, RY and JC contributed to the Investigation. FT, RY, JC, MC, MR and GEHF contributed to Methodology development. FT, AC and GEHF contributed to Project Administration. FT and JC contributed to Software development. FT contributed to Validation. FT and RY contributed to Visualization. RY, MC, MR and GEHF contributed to Data Curation. FT and RY contributed to Writing of the original draft. FT, RY, MC, GEHF and AC contributed to the Review & Editing of subsequent drafts. GEHF contributed to Funding Acquisition and Resource Provision. GEHF and AC provided Supervision of the project.

Funding

None.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 12 December 2024 / Accepted: 29 April 2025
Published online: 10 May 2025

References

1. 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. https://doi.org/10.1136/bmjopen-2016-012545.
2. Bornmann L, Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. J Assoc Inf Sci Technol. 2015;66(11):2215–22. https://doi.org/10.1002/asi.23329.
3. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev. 2015;4(1):80. https://doi.org/10.1186/s13643-015-0067-6.
4. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210. https://doi.org/10.1186/s13643-016-0384-4.
5. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of distillersr's machine learning-based prioritization tool for title/abstract screening– impact on reviewer-relevant outcomes. BMC Med Res Methodol. 2020;20(1):256. https://doi.org/10.1186/s12874-020-01129-1.
6. van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021;3(2):125–33. https://doi.org/10.1038/s42256-020-00287-7.
7. Cheng Sh, Augustin C, Bethel A, et al. Using machine learning to advance synthesis and use of conservation and environmental evidence. Conserv Biol. 2018;32(4):762–4. https://doi.org/10.1111/cobi.13117.
8. Harrison H, Griffin SJ, Kuhn I, Usher-Smith JA. Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. BMC Med Res Methodol. 2020;20(1):7. https://doi.org/10.1186/s12874-020-0897-3.
9. Issaiy M, Ghanaati H, Kolahi S, et al. Methodological insights into ChatGPT's screening performance in systematic reviews. BMC Med Res Methodol. 2024;24(1):78. https://doi.org/10.1186/s12874-024-02203-8.
10. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large Language models: data analysis study. J Med Internet Res. 2024;26(1):e48996. https://doi.org/10.2196/48996.
11. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. 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. Res Synth Methods. 2024;15(4):616–26. https://doi.org/10.1002/jrsm.1715.
12. Chakhtoura M, Nassar JE, Slim A et al. Vitamin D Supplementation and Falls in Adults: A Systematic Umbrella Review of Meta-Analyses of Randomized Controlled Trials. In: *Annual Meeting of the Endocrine Society, Boston, June 1–4*, 2024. SUN-268.
13. Chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis. Accessed May 23. 2024. https://training.cochrane.org/handbook/current/chapter-03
14. Valizadeh A, Moassefi M, Nakhostin-Ansari A, et al. Abstract screening using the automated tool Rayyan: results of effectiveness in three diagnostic test accuracy systematic reviews. BMC Med Res Methodol. 2022;22(1):160. https://doi.org/10.1186/s12874-022-01631-8.

15. Li J, Kabouji J, Bouhadoun S, et al. Sensitivity and specificity of alternative screening methods for systematic reviews using text mining tools. J Clin Epidemiol. 2023;162:72–80. https://doi.org/10.1016/j.jclinepi.2023.07.010.
16. Dos Reis AHS, De Oliveira ALM, Fritsch C, Zouch J, Ferreira P, Polese JC. Usefulness of machine learning softwares to screen titles of systematic reviews: a methodological study. Syst Rev. 2023;12(1):68. https://doi.org/10.1186/s13643-023-02231-3.
17. Luo X, Chen F, Zhu D, et al. Potential roles of large Language models in the production of systematic reviews and Meta-Analyses. J Med Internet Res. 2024;26:e56780. https://doi.org/10.2196/56780.
18. Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial intelligence to Automate network Meta-Analyses: four case studies to evaluate the potential application of large Language models. PharmacoEconomics - Open. 2024;8(2):205–20. https://doi.org/10.1007/s41669-024-00476-9.
19. Nashwan AJ, Jaradat JH. Streamlining systematic reviews: Harnessing large Language models for quality assessment and Risk-of-Bias evaluation. Cureus Published Online August. 2023;6. https://doi.org/10.7759/cureus.43023.
20. Hasan B, Saadi S, Rajjoub NS et al. Integrating large Language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. BMJ Evid-Based Med. Published online February 21, 2024:bmjebm–2023. https://doi.org/10.1136/bmjebm-2023-112597
21. Barsby J, Hume S, Lemmey HA, Cutteridge J, Lee R, Bera KD. Pilot study on large Language models for risk-of-bias assessments in systematic reviews: A(I) new type of bias? *BMJ Evid-Based med*. Published online May 23, 2024:bmjebm-2024-112990. https://doi.org/10.1136/bmjebm-2024-112990
22. Huotala A, Kuutila M, Ralph P, Mäntylä M. The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews. In: *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*. EASE '24. Association for Computing Machinery; 2024:262–271. https://doi.org/10.1145/3661167.3661172
23. Matsui K, Utsumi T, Aoki Y, Maruki T, Takeshima M, Takaesu Y. Human-Comparable sensitivity of large Language models in identifying eligible studies through title and abstract screening: 3-Layer strategy using GPT-3.5 and GPT-4 for systematic reviews. J Med Internet Res. 2024;26:e52758. https://doi.org/10.2196/52758.
24. Li M, Sun J, Tan X. Evaluating the effectiveness of large Language models in abstract screening: a comparative analysis. Syst Rev. 2024;13(1):219. https://doi.org/10.1186/s13643-024-02609-x.
25. Oami T, Okada Y, Nakada T, aki. Performance of a large Language model in screening citations. JAMA Netw Open. 2024;7(7):e2420496. https://doi.org/10.1001/jamanetworkopen.2024.20496.
26. Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. Title and abstract screening for literature reviews using large Language models: an exploratory study in the biomedical domain. Syst Reviews. 2024;13(1):158.

Publisher's note