Evaluating Large Language Models for Literature Screening: A Systematic Review of Sensitivity and Workload Reduction

Elias Sandner[1,5][a], Luca Fontana[2][b], Kavita Kothari[3][c], Andre Henriques[4][d], Igor Jakovljevic[1][e], Alice Simniceanu[2][f], Andreas Wagner[1][g] and Christian Gütl[5][h]

[1] *IT Department, CERN, Geneva, Switzerland*

[2] *Health Emergencies Programme, WHO, Geneva, Switzerland*

[3] *Consultant to Library & Digital Information Networks, WHO, Kobe, Japan*

[4] *Occupational Health & Safety and Environmental Protection (HSE) Unit, CERN, Geneva, Switzerland*

[5] *Cognitive & Digital Science Lab, Technical Unversity Graz, Graz, Austria*

Keywords: Systematic Review, Evidence Synthesis, Large Language Models, Literature Screening Automation, Binary Text Classification.

Abstract: Systematic reviews provide high-quality evidence but require extensive manual screening, making them time-consuming and costly. Recent advancements in general-purpose large language models (LLMs) have shown potential for automating this process. Unlike traditional machine learning, LLMs can classify studies based on natural language instructions without task-specific training data. This systematic review examines existing approaches that apply LLMs to automate the screening phase. Models used, prompting strategies, and evaluation datasets are analyzed, and the reported performance is compared in terms of sensitivity and workload reduction. While several approaches achieve sensitivity above 95%, none consistently reach the 99% threshold required for replacing human screening. The most effective models use ensemble strategies, calibration techniques, or advanced prompting rather than relying solely on the latest LLMs. However, generalizability remains uncertain due to dataset limitations and the absence of standardized benchmarking. Key challenges in optimizing sensitivity are discussed, and the need for a comprehensive benchmark to enable direct comparison is emphasized. This review provides an overview of LLM-based screening automation, identifying gaps and outlining future directions for improving reliability and applicability in evidence synthesis.

1 INTRODUCTION

By synthesizing findings from potentially all relevant studies on a given research question, a Systematic Review (SR) represents the most reliable research methodology for evidence-based conclusions (Shekelle et al., 2013). Therefore, SRs play a crucial role in the medical field, guiding decision-making and shaping clinical practice guidelines (Cook et al., 1997). However, the rigor of systematic reviews makes them highly time- and resource-intensive, often taking months or even years to complete.

Systematic reviews typically begin with a broad database query to ensure comprehensive coverage, followed by human screening—a particularly time-consuming stage of the process (Carver et al., 2013).

Despite following a well-defined procedure, automating the screening phase remains challenging. Existing methods often fall short of human-level sensitivity and lack generalizability across review domains. Traditional ML approaches can support large-scale or living SRs, but their effectiveness is limited by the scarcity of high-quality training data. (Sandner et al., 2024a)

General-purpose LLMs have shown strong performance in classification tasks. Trained on vast text

[a] https://orcid.org/0009-0007-9855-4923
[b] https://orcid.org/0000-0002-8614-4114
[c] https://orcid.org/0000-0002-0759-5225
[d] https://orcid.org/0000-0003-1521-3423
[e] https://orcid.org/0000-0003-1893-9553
[f] https://orcid.org/0000-0003-4068-6177
[g] https://orcid.org/0000-0001-9589-2635
[h] https://orcid.org/0000-0001-9589-1966

Sandner, E., Fontana, L., Kothari, K., Henriques, A., Jakovljevic, I., Simniceanu, A., Wagner, A., Gütl and C.
Evaluating Large Language Models for Literature Screening: A Systematic Review of Sensitivity and Workload Reduction.
DOI: 10.5220/0013562900003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 508-517
ISBN: 978-989-758-758-0; ISSN: 2184-285X

corpora, they exhibit human-like reasoning and can follow natural language instructions to perform classification without task-specific training. (Zhou et al., 2024; Carneros-Prado et al., 2023)

For literature screening, eligibility criteria combined with a study's title and abstract are used as input to an LLM-based framework, which classifies studies as included or excluded, emulating human decision-making.

The key requirement for integrating such tools into the workflow is minimizing the risk of wrongly excluding relevant studies, measured by sensitivity. While some studies accept 95% sensitivity (Bramer et al., 2017; Callaghan and Müller-Hansen, 2020), Cochrane[1]—a leading authority in high-quality SRs—requires 99% sensitivity for tools replacing human screening (Thomas et al., 2021).

Despite progress in automating literature screening, fully replacing human screeners remains unlikely in the near future. Until then, such systems can be used to pre-filter studies and reduce researchers' workload—measured by the number of excluded records, which should be maximized. When balancing sensitivity and workload reduction, low sensitivity makes a system unsuitable due to the risk of missing relevant studies. In contrast, any workload reduction improves upon manual screening—making sensitivity the top priority.

Previous research showed promising results with a 5-tier prompting approach, theoretically applicable to any SR, though its generalizability is limited due to the specific reviews used for evaluation (Sandner et al., 2024b). During this case study, it also became evident that the literature lacks a comprehensive overview of similar methods. This SR addresses that gap by reviewing the most promising applications of general-purpose LLMs for literature screening in evidence synthesis. It summarizes the models, prompts, and evaluation datasets, compares performance in terms of sensitivity and workload reduction, and presents additional metrics in the supplementary material. The review addresses the following research question: *Which studies have investigated the use of general-purpose LLMs to automate the screening process in systematic literature reviews, and what insights can be drawn from the most effective approaches?*

[1] https://www.cochrane.org/

2 METHODOLOGY

The methodology of this SR builds on principles outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021) and the Cochrane Handbook for Systematic Reviews of Interventions (Higgins et al., 2024), adapted to suit the context of computer science research. In addition, the methodology was guided by insights from Carrera-Rivera et al. (2022)'s guide on conducting a SR in the domain of computer science research.

2.1 Study Identification

The methodology begins with retrieving relevant studies from the following academic databases: Europe PMC (Europe PMC, 2025), Web of Science (Clarivate, 2025), Embase (Elsevier, 2025), and Medline-OVID (Wolters Kluwer, 2025). An information specialist on the author team developed tailored search strategies for each database through an iterative process, using seed papers to ensure relevance. All search strategies are available in the supplementary material[2].

All searches were executed on June 10, 2024. Retrieved studies underwent deduplication through Covidence[3]'s built-in feature.

2.2 Study Selection

Inclusion and exclusion criteria were defined using the PICO (Population, Intervention, Comparison, Outcomes) framework, as recommended in the considered guidelines. English-language studies from 2022 onward were included, while editorials, commentaries, and book chapters were excluded. Eligible studies investigated the use of general-purpose LLMs for the screening phase of systematic reviews at either the title-abstract (TiAb) or full-text level. Studies were excluded if they employed specialized LLMs (e.g., fine-tuned for review-specific classification tasks), traditional statistical classifiers, or decision-support systems requiring human intervention for final decisions. Studies were considered if they compared LLM-based decisions to human screening judgments, either retrospectively or based on data labeled within the study. Exclusion also applied to studies that did not report sensitivity or workload reduction and lacked sufficient information to calculate these metrics, or failed to disclose the dataset used for evaluation.

[2] https://zenodo.org/records/15255994

[3] https://www.covidence.org/

Title and abstract screening was conducted using the free version of Covidence, while the free version of Rayyan[4] was utilized for full-text screening. Both screening phases were performed independently and in duplicate by two human reviewers. In both phases, conflicts were resolved through discussion between the two reviewers.

2.3 Data Extraction and Analysis

Following the full-text screening, articles that met the eligibility criteria were subject to data extraction, executed by one author using a spreadsheet tool and a pre-developed extraction sheet.

For citations describing multiple experiments with varied models or prompts, the focus was placed on the approach reporting the highest sensitivity. Supplementary experiments were considered if they provided meaningful insights for comparison with the main approach or exhibited significant differences from it.

For each considered experiment, the model used, as well as detailed information on the prompt, dataset, and reported performance were extracted. The applied prompting strategy was recorded, along with the characteristics exhibited by the prompt Furthermore, the parameters inserted into the prompt template and the expected response from the LLM, based on the instructions provided in the prompt, were extracted.

Literature screening automation is typically evaluated on labeled bibliographic records. Extracted dataset characteristics include the number of reviews, total records, and records labeled as 'include'. Additional details include the screening stage at which labels were assigned (title/abstract or full text), whether labels reflected a blinded consensus by two reviewers, as well as the dataset domain and public availability.

Performance-related data were also extracted. Since outcomes are often presented in tables using diverse metrics, full tables were collected initially. In a subsequent step, sensitivity and workload reduction were extracted or calculated. These two parameters are reported in this SR based on the following definitions:

Sensitivity, as defined in (1), refers to the ability of the screening system to correctly identify all relevant studies. It measures the portion of actual positives (relevant studies) that are correctly identified as such by the system and is crucial in the given context as it measures the risk of missing relevant literature.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (1)$$

[4]https://www.rayyan.ai/

Assuming that the LLM-based screening automation is integrated into the SR workflow as a filtration step, human experts subsequently have to screen those records classified as include while those classified as exclude are no longer subject to the time-consuming manual screening task. Consequently, the **workload reduction (WR)** as defined in (2) is the fraction of papers excluded by the model.

$$\text{WR} = \frac{\text{True Negative} + \text{False Negative}}{\text{N}} \quad (2)$$

where N represents the total number of papers.

3 RESULTS

This chapter presents the outcomes of the SR. It begins with the study selection process and the identified approaches. Then, it describes how sensitivity and workload reduction were extracted. Finally, it compares the selected screening automation solutions by methodology, outlines the evaluation datasets, and summarizes the results. Additionally, the supplementary material[5] provides comprehensive details, including the complete extracted data and links to the datasets used in the cited studies.

3.1 Selection of Screening Automation Approaches

The study selection process is depicted in the Figure 1. Out of 280 unique retrieved studies, 256 have been excluded in the TiAb screening phase. Out of the remaining 24 papers, 19 were retrieved as full text. After full-text screening 12 studies turned out to fulfill the defined eligibility criteria and have therefore been subject of data extraction. For one of the 12 publications, we identified a numerical inconsistency which resulted in excluding the paper as detailed in 3.2.

All selected papers proposed approaches for automating SR screening with general-purpose LLMs and benchmarked their performance against human decisions.

Most papers did not only describe one experiment but compared multiple prompting strategies, models, or datasets, reporting results separately. In this SR, each study is represented by the approach with the highest sensitivity. If the selected approach was tested on several datasets, efforts were made to calculate the average result across all datasets.

To account for methodological variations, results from three studies were reported with multiple cases

[5]https://zenodo.org/records/15255994

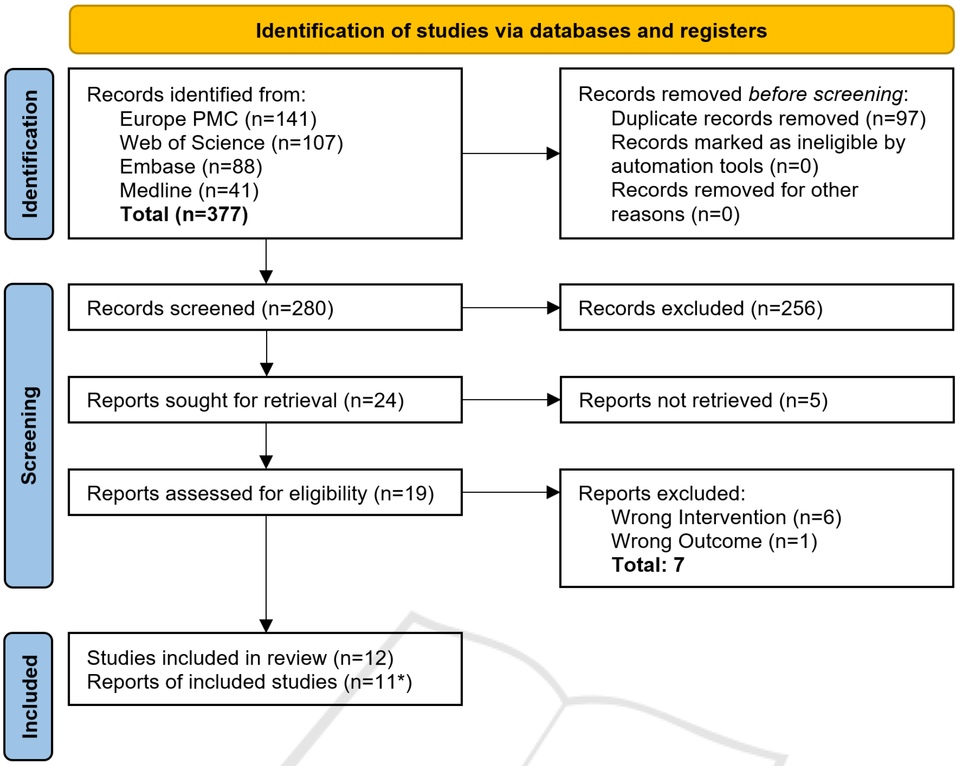

Figure 1: PRISMA flowchart for the identification and selection of studies according to (Page et al., 2021); *As outlined in 3.2, one of the 12 selected study was subsequently excluded due to inconsistencies in the reported numbers.

where different experimental setups provided valuable insights for comparison. Cai et al. (2023) reports both a single-shot approach and two multi-shot approaches. To allow for a direct comparison, the single-shot approach and the better-performing multi-shot approach were selected. Akinseloyin et al. (2024) models the screening task as a relevance ranking problem, where only the top k% papers are retained for human review. For this study, results based on two different threshold settings were included to reflect variations in the ranking-based approach. Cao et al. (2024) evaluated seven prompt strategies using title and abstract information, and six using full-text. To represent both categories, the best-performing strategy from each was included. Consequently, the subsequent sections summaries and discusses 14 approaches out of 8 publications and 3 preprints.

3.2 Mathematical Inference

While sensitivity is typically reported directly, workload reduction is discussed in most papers but defined inconsistently. Furthermore, several papers did not report these metrics across all considered datasets. Therefore, additional calculations were required beyond the standard data extraction procedure.

Fortunately, in addition to sensitivity, performance metrics such as specificity, accuracy, precision, F1-score, F3-score, and positive/negative predictive values were reported. Combined with the total number of records and the number of ground-truth inclusions, these metrics enabled mathematical inference to determine true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From these values, sensitivity and workload reduction across all considered data records were derived.

The numerical values required for these calculations were either directly reported in the publications or obtained from supplementary materials (e.g., data, code, documentation) and references describing the datasets used. Additionally, when necessary, the authors were contacted to provide further information. These inferences were made in accordance with the reported data and to the best of our knowledge, with all details transparently documented in the supplementary material[6].

For one of the 12 publications, a numerical inconsistency in these calculations could not be resolved. Consequently, despite meeting the predefined eligibility criteria, this publication was retrospectively ex-

[6]https://zenodo.org/records/15255994

cluded.

3.3 Models and Prompts

Table 1 presents an overview of the models used and outlines the prompting strategy applied in selected screening automation approaches, which is subsequently detailed.

While several selected studies tested multiple LLMs, the best performance was reported with GPT-3.5-turbo in 6 out of 11. The approach described in Guo et al. (2024) switched to GPT-4 when the context length of GPT-3.5-turbo was exceeded. Notably, none of the papers favoring GPT-3.5-turbo compared it to GPT-4. Furthermore, four papers reported best results by utilizing GPT-4, the most advanced OpenAI model at the time of search execution.

In two studies, the best results were achieved using ensemble models that combined the results of more than one LLM. Li et al. (2024) employed Latent Class Analysis (LCA) (McCutcheon, 1987) based on responses from GPT-4, GPT-3.5, and LLaMA-2 to determine the screening decisions. Wang et al. (2024) utilized two LLaMA-2 models (7b-ins and 13b-ins) along with the language model BioBERT. The model outcomes were fused using CombSUM (Fox and Shaw, 1994).

All applied prompting approaches instruct the LLM to screen one specific study at a time. While eight follow a single-shot approach, five split the task into more than one prompt, utilizing a multi-prompt approach.

The approach described in Tran et al. (2023) requires the eligibility criteria to be provided in PICOS format. For each PICOS category (Population, Intervention, Comparison, Outcome, Study Design) an individual request is sent to the LLM. Similarly, Cai et al. (2023) sends individual requests for each eligibility criterion. In both approaches, a record is excluded if any of the criteria are violated.

Spillias et al. (2024) executed three repeated calls to the same LLM, each complemented by a random context string. The final decision was made based on a voting strategy. It was reported that this approach improved the quality of screening beyond what could be achieved by optimizing OpenAI's temperature parameter.

A multi-prompt approach to enable efficient full-text screening was applied by Khraisha et al. (2024). The full text is divided into segments, which are subsequently provided to the LLM, and the process stops if all criteria are met.

Akinseloyin et al. (2024) introduced a framework for screening automation that first utilizes the LLM to transform eligibility criteria into multiple yes/no questions. Each question is then sent in a separate prompt, expecting a free-text response. The sentiment of these responses is analyzed using a BART model, resulting in a likelihood score of the response being positive. Additionally, the cosine similarity between the question and the abstract is computed. The final question score is calculated by averaging the sentiment score with the cosine similarity. To calculate the paper's final score, the average of all question scores is first computed, which is then further averaged with the cosine similarity between all eligibility criteria and the abstract. Finally, all studies are ranked in descending order based on their final score, with the top k% classified as "include" and the rest as "exclude."

As the LLM is instructed to assess a record's relevance to an SR, corresponding information must be provided in the prompt. Typically, human screeners base their decisions on the title, abstract, and eligibility criteria, which was reflected in most approaches. Wang et al. (2024) provided only the review title but no criteria, while others included the review title, topic, or objective in addition to the criteria. Khraisha et al. (2024) and Cao et al. (2024) (ISO-Screen-Prompt) also considered the full text to inform the LLM's decision.

Although all frameworks output a binary classification ('included' or 'excluded'), they differ in the expected LLM response format. Nine approaches prompt the LLM to reply with one of two specified keywords. Of these, two require additional reasoning insights in the response. Li et al. (2024) expects the LLM to return a binary decision for each criterion along with reasoning for the decision. Spillias et al. (2024) expects the LLM to reason about the initial decision, reflect on it, make a final decision, and then provide reasoning again. Cai et al. (2023) allows the LLM to respond with one of three terms: 'yes,' 'no,' or 'not sure.' To increase sensitivity, responses of 'not sure' are considered as 'include' decisions. Issaiy et al. (2024) expects the LLM to respond with a rating from one to five. Subsequently, papers rated as three to five are treated as 'include' decisions, while those rated one or two are considered 'exclude' decisions.

Prompt phrasing significantly influences the model's reasoning and decisions. Many prompts use roleplay, casting the LLM as a researcher or reviewer to simulate human judgment. Others apply Chain of Thought (CoT) prompting, guiding the model through reasoning steps before a final decision. Both techniques aim to enhance performance through deeper, more consistent reasoning. Notably, Cao et al. (2024)'s ISO-Screen Prompt uses 'instruction repeti-

Table 1: Summary of Employed Models and Prompting Strategies, Including Key Characteristics of Screening Automation Methods: References in bold indicate approaches that achieve a sensitivity above 95%; Differences between approaches originating from the same study are underlined.

Reference	Model	Prompt Strategy	Prompt Parameters	Prompt Return	Prompt Characteristic
(Wang et al., 2024)	Ensamble (LLaMA2-7b-ins, LLaMA2-13b-ins, BioBERT)	Single-Shot	**Review Title**, Title, Abstract	Binary, Extracted LLM Confidence Score	Instruction based
(Cao et al., 2024) - Abstract Screen Prompt	GPT-4	Single-Shot	Review Objectives, Title, Abstract, Eligibility Criteria	Binary	Roleplay
(Cao et al., 2024) - ISO Screen Prompt	GPT-4	Single-Shot	Review Objectives, Eligibility Criteria, full-text	Binary	Roleplay, Repetition of Instruction
(Akinseloyin et al., 2024)	GPT-3.5-Turbo	Multi-Shot	Question Generation: Eligibility Criteria Question Answering: Review Title, Abstract, Question	Question Generation: 5 Yes/No Questions Question Answering: Answer of Question in Natural Text	Roleplay
(Issaiy et al., 2024)	GPT-3.5-Turbo	Single-Shot	Title, Abstract, Reference Type, Date, Eligibility Criteria (PICOS)	Score from 1 to 5	Instruction Based
(Li et al., 2024)	Ensamble (GPT-4, GPT-3.5, LLaMA-2)	Single-Shot	Review Topic, Title, Abstract, Eligibility Criteria	Binary (Final Decision), Binary (for each Criteria), Overall Reasoning	Chain of Thought
(Tran et al., 2023)	GPT-3.5-Turbo	Multi-Shot	Title, Abstract, Eligibility Criteria (PICOS)	Binary	Chain of Thought
(Spillias et al., 2024)	GPT-3.5-Turbo	Multi-Shot	Random String, Title, Abstract, Eligibility Criteria	Binary (Initial Decision), Reasoning, Reflection, Binary (Final Decision), Reasoning	Roleplay, Chain of Thought, Random String
(Guo et al., 2024)	GPT-3.5-Turbo*	Single-Shot	Title, Abstract, Eligibility Criteria	Binary	Roleplay
(Gargari et al., 2024)	GPT-3.5-Turbo	Single-Shot	Review Title, Title, Abstract, Eligibility Criteria	Binary	Instruction Based
(Khraisha et al., 2024)	GPT-4	Multi-Shot	Full-Text Segment, Eligibility Criteria	Binary	Roleplay, extensive instruction on eligibility critera
(Cai et al., 2023) - Instruction Prompt	GPT-4	Single-Shot	Title, Abstract, Eligibility Criteria	Binary	Roleplay
(Cai et al., 2023) - Single Criterion	GPT-4	Multi-Shot	Title, Abstract, Eligibility Criteria	Binary + ”not sure”, Reasoning	Roleplay

tion,' placing the task description before and after the full text—likely reinforcing focus and improving performance.

3.4 Evaluation and Performance Comparison

Each screening automation approach was evaluated by benchmarking against human screening decisions. Therefore, labeled datasets were considered as ground truth and compared with the final binary decision of each approach. Table 2 describes the datasets on which the given approaches were tested and reports their classification performance.

The size and variety of the used datasets indicate the generalizability of the reported results. Applied datasets focus on different areas within the medical domain ranging from pharmacology intervention studies to social health qualitative studies. The only exception is the dataset used by Spillias et al. (2024), which covers data from a single SR on Community-Based Fisheries Management. This dataset is also the only one where the ground truth annotation was executed by a single screener, whereas all other datasets are based on double-blind screening or annotations from even more human reviewers.

Only two datasets consist of data from more than 10 SRs, and only five encompass more than 10,000 records. Especially noteworthy are the datasets used by (Wang et al., 2024), who conducted experiments on datasets released as part of CLEF TAR from 2017 to 2019 (Kanoulas et al., 2017, 2018, 2019). Together, these datasets contain data from more than 100 SRs, encompassing over 600,000 records. When interpreting these numbers and the associated performance, note that Wang et al. (2024) applied a leave-one-out calibration approach. In other words, data from all but one review were used to calibrate the threshold, and the remaining review was used for validation. Consequently, the threshold was fine-tuned for each review rather than determined universally. Nevertheless, the approach was evaluated on the complete dataset.

Performance varies significantly, even though the approaches follow similar principles. Six approaches reported a sensitivity above 95%, which is a commonly applied target (Bramer et al., 2017; Callaghan and Müller-Hansen, 2020). As advised by the Cochrane Information Retrieval Methods Group (IRMG), systems designed to reduce the manual screening workload for high-quality SRs must be calibrated to a sensitivity greater than 99% to replace human screening (Thomas et al., 2021). However, none of the considered approaches consistently reached this value across the considered datasets. Workload reduction, defined as the fraction of papers excluded by the system and consequently not requiring human screening, varied from 29% to almost 100%. Approaches that achieved a sensitivity above 95% reached workload reductions ranging from 48% to 79%.

4 DISCUSSION

This section aims to highlight the factors underlying strong outcomes and identify commonalities among the studies that contributed to the field by reporting on approaches that lacked sufficient sensitivity. Given that the experiments across the selected studies were evaluated on different datasets, direct comparisons are not possible, and any conclusions drawn in this section require further validation. For comparisons of the reported approaches with similar ones evaluated on the same dataset, please refer to the cited papers.

While the suitability of approaches with a sensitivity of 95% for replacing human screening remains a topic of discussion and highly depends on the use case, lower sensitivities are widely considered insufficient. In this context, both approaches utilizing an ensemble model (Wang et al., 2024; Li et al., 2024) and those incorporating calibration, either based on next-token likelihood (Wang et al., 2024) or by expecting the model to provide a score (Issaiy et al., 2024), have been observed to achieve this threshold. The framework introduced by Akinseloyin et al. (2024) demonstrated that incorporating similarity scores improves performance which resulted in achieved this threshold as well. Experiments conducted by Cao et al. (2024) may have achieved their strong results due to the use of GPT-4, the most advanced model among those considered, combined with exhaustive prompt engineering. Interestingly, their experiments also suggest that incorporating the full text does not lead to further improvements in performance. To further increase sensitivity toward meeting Cochrane's requirement of 99% (Thomas et al., 2021), combining these approaches is a promising direction for future work.

Furthermore, it is noteworthy that the approach by Wang et al. (2024), which utilized LLaMA-2 models instead of more advanced ones and employed a relatively simple prompt design, achieved the highest sensitivity along with a substantial workload reduction of 72%. Considering that the prompt did not include any eligibility criteria and the LLM assessed record relevance solely based on the review title, it can be hypothesized that eligibility criteria, designed to guide human screeners, may be interpreted by LLMs either too strictly or as too complex to process effectively.

Table 2: Characteristics of Evaluation Datasets and Reported Performance of included screening automation approaches: Ground truth column indicates wether human annotation from the Title and Abstract (TiAb) or the full-text (FT) screening phase were considered as gold-standard; The table is Sorted by decreasing sensitivity.

	Dataset				Results	
Reference	**No. of Reviews**	**No. of Records**	**No. of Includes**	**Ground Truth**	**Sensitivity**	**Workload Reduction**
(Wang et al., 2024)	128	657,980	10,524	TiAb	**97%**	**72%**
(Cao et al., 2024) - Abstract Screen Prompt	10	4000	779	TiAb	**97%**	**70%**
(Cao et al., 2024) - ISO Screen Prompt	10	3230	487	TiAb	**96%**	**79%**
(Akinseloyin et al., 2024) - top 50%	31	76,025	1710	TiAb	**96%**	**50%**
(Issaiy et al., 2024)	6	1180	148	TiAb	**95%**	**48%**
(Li et al., 2024)	3	505	205	FT	**95%**	**60%**
(Tran et al., 2023)	5	22,666	1485	TiAb	91%	29%
(Spillias et al., 2024)	1	1098	101	TiAb	85%	88%
(Akinseloyin et al., 2024) - top 20%	31	76,025	1710	TiAb	80%	80%
(Guo et al., 2024)	6	24,845	538	TiAb	76%	90%
(Gargari et al., 2024)	1	330	13	FT	62%	99%
(Khraisha et al., 2024)	1	150	39	FT	57%	73%
(Cai et al., 2023) - Instruction Prompt	4	400	40	TiAb	51%	79%
(Cai et al., 2023) - Single Criterion	4	400	40	TiAb	41%	89%

As a result, this misinterpretation may lead to incorrect exclusions. Therefore, further analysis on how to effectively instruct the LLM and determine which information it should consider when making inclusion/exclusion decisions would be a highly relevant contribution for future work.

The calibration approach applied by Wang et al. (2024) was based on reviews that closely resembled the one used for evaluation, and similar performance might not be achieved in use cases with less similar reviews. In theory, applying this approach on a more general dataset should enable similar sensitivity due to the calibration. However, this may come at the cost of a significant decrease in workload reduction.

From studies that resulted in lower sensitivity, it can be concluded that solely relying on a model with high reasoning capabilities, such as GPT-4, is not sufficient. Furthermore, evaluating candidate studies for each criterion separately does not necessarily improve sensitivity, nor does providing full texts as segmented inputs in subsequent prompts. Therefore, while multi-shot approaches significantly increase computing costs, they offer no notable advantages. However, lower sensitivity may also be influenced by factors such as the specific dataset used and the complexity of the underlying SRs.

Considering that most studies conducted their evaluations on a relatively small number of records, with data originating from a limited set of reviews or highly similar reviews, it is difficult to argue that similar sensitivity would be achieved in real-world applications. However, to the best of our knowledge, no clear guidelines have been established for evaluating screening automation solutions to determine their trustworthiness as a replacement for human screening. To enable direct comparison and gain trust from the evidence synthesis community, a standardized benchmark should be established, along with clear requirements for performance evaluation. This benchmark should not be restricted to a specific type of SR and should be as extensive as possible in terms of both the number of SRs contributing data and the number of candidate studies included. The datasets used in the studies considered in this review could serve as a valid foundation for developing such a benchmark.

5 CONCLUSION

This SR provides a comprehensive overview of existing approaches that leverage general-purpose LLMs for automating literature screening in evidence synthesis. By summarizing models, prompts, and evaluation datasets, as well as comparing their sensitivity and workload reduction, this review highlights key trends and challenges in the field.

The findings indicate that achieving high sensitivity remains a primary challenge, particularly given Cochrane's recommended threshold of 99% for reliable automation. While some approaches, such as ensemble models and those incorporating calibration mechanisms, reached sensitivity levels above 95%, no single method consistently met the highest standard. Notably, the approach resulting in the highest sensitivity utilized LLaMA-2 models combined with a rather simple prompt design, demonstrating that complex solutions not always be necessary for strong performance. However, the generalizability of presented results remains uncertain, as evaluation was conducted on either relatively small datasets or after fine-tuning based on highly similar reviews.

Additionally, findings suggest that solely relying on advanced reasoning capabilities of models like GPT-4, segmenting full texts, or evaluating each eligibility criterion separately does not necessarily enhance sensitivity. Instead, future research should explore combining effective techniques, optimizing prompt design, and expanding dataset diversity to improve performance.

A key limitation in current research is the absence of a standardized benchmark for evaluating screening automation, which complicates the assessment of effectiveness. Establishing a benchmark with well-defined performance criteria is essential to enhance transparency and credibility within the evidence synthesis community. This benchmark should incorporate a diverse set of SRs and large datasets to enable rigorous and reproducible comparisons across different approaches. The datasets analyzed in this review could serve as a foundation for such an initiative.

ACKNOWLEDGEMENTS

The joint CERN and WHO ARIA[7] project is funding the PhD project, in the context of which this systematic review was conducted.

In (Sandner et al., 2024b), we described the 5-tier prompting approach as novel. In the context of this systematic review it was discovered that (Issaiy et al., 2024) describes a very similar approach. Therefore, we acknowledge them as the first one introducing the strategy of classifying into more than one category and subsequently transforming the result into a binary format to calibrate the system towards higher sensitivity.

REFERENCES

Akinseloyin, O., Jiang, X., and Palade, V. (2024). A question-answering framework for automated abstract screening using large language models. *Journal of the American Medical Informatics Association*, 31(9):1939–1952.

Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., and Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Systematic reviews*, 6:1–12.

Cai, X., Geng, Y., Du, Y., Westerman, B., Wang, D., Ma, C., and Vallejo, J. J. G. (2023). Utilizing chatgpt to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. *medRxiv*, pages 2023–09.

Callaghan, M. W. and Müller-Hansen, F. (2020). Statistical stopping criteria for automated screening in systematic reviews. *Systematic Reviews*, 9:1–14.

Cao, C., Sang, J., Arora, R., Kloosterman, R., Cecere, M., Gorla, J., Saleh, R., Chen, D., Drennan, I., Teja, B., et al. (2024). Prompting is all you need: Llms for systematic review screening. *medRxiv*, pages 2024–06.

Carneros-Prado, D., Villa, L., Johnson, E., Dobrescu, C. C., Barragán, A., and García-Martínez, B. (2023). Comparative study of large language models as emotion and sentiment analysis systems: A case-specific analysis of gpt vs. ibm watson. In *International Conference on Ubiquitous Computing and Ambient Intelligence*, pages 229–239. Springer.

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., and Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9:101895.

Carver, J. C., Hassler, E., Hernandes, E., and Kraft, N. A. (2013). Identifying barriers to the systematic literature review process. In *2013 ACM/IEEE international symposium on empirical software engineering and measurement*, pages 203–212. IEEE.

[7]https://partnersplatform.who.int/tools/aria

Clarivate (2025). Web of science: Advanced search. Accessed: February 27, 2025.

Cook, D. J., Greengold, N. L., Ellrodt, A. G., and Weingarten, S. R. (1997). The relation between systematic reviews and practice guidelines. *Annals of internal medicine*, 127(3):210–216.

Elsevier (2025). Embase: Advanced search. Accessed: February 27, 2025.

Europe PMC (2025). Europe pmc: An archive of life sciences literature. Accessed: February 27, 2025.

Fox, E. and Shaw, J. (1994). Combination of multiple searches. *NIST special publication SP*, pages 243–243.

Gargari, O. K., Mahmoudi, M. H., Hajisafarali, M., and Samiee, R. (2024). Enhancing title and abstract screening for systematic reviews with gpt-3.5 turbo. *BMJ Evidence-Based Medicine*, 29(1):69–70.

Guo, E., Gupta, M., Deng, J., Park, Y.-J., Paget, M., and Naugler, C. (2024). Automated paper screening for clinical reviews using large language models: Data analysis study. *Journal of Medical Internet Research*, 26:e48996.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A., editors (2024). *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane, version 6.5 (updated august 2024) edition.

Issaiy, M., Ghanaati, H., Kolahi, S., Shakiba, M., Jalali, A. H., Zarei, D., Kazemian, S., Avanaki, M. A., and Firouznia, K. (2024). Methodological insights into chatgpt's screening performance in systematic reviews. *BMC Medical Research Methodology*, 24(1):78.

Kanoulas, E., Li, D., Azzopardi, L., and Spijker, R. (2017). Clef 2017 technologically assisted reviews in empirical medicine overview. In *CEUR Workshop Proceedings*. CEUR-WS.org.

Kanoulas, E., Li, D., Azzopardi, L., and Spijker, R. (2018). Clef 2018 technologically assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, volume 2125.

Kanoulas, E., Li, D., Azzopardi, L., and Spijker, R. (2019). Clef 2019 technology assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, volume 2380, page 250.

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., and Hadfield, K. (2024). 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*.

Li, M., Sun, J., and Tan, X. (2024). Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. *Systematic reviews*, 13(1):219.

McCutcheon, A. L. (1987). *Latent class analysis*. Sage.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Sandner, E., Gütl, C., Jakovljevic, I., and Wagner, A. (2024a). Screening automation in systematic reviews: Analysis of tools and their machine learning capabilities. In *dHealth 2024*, pages 179–185. IOS Press.

Sandner, E., Hu, B., Simiceanu, A., Fontana, L., Jakovljevic, I., Henriques, A., Wagner, A., and Gütl, C. (2024b). Screening automation for systematic reviews: A 5-tier prompting approach meeting cochrane's sensitivity requirement. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 150–159. IEEE.

Shekelle, P. G., Maglione, M. A., Luoto, J., et al. (2013). *Global Health Evidence Evaluation Framework*. Agency for Healthcare Research and Quality (US), Rockville, MD. Table B.9, NHMRC Evidence Hierarchy: designations of 'levels of evidence' according to type of research question (including explanatory notes).

Spillias, S., Tuohy, P., Andreotta, M., Annand-Jones, R., Boschetti, F., Cvitanovic, C., Duggan, J., Fulton, E. A., Karcher, D. B., Paris, C., et al. (2024). Human-ai collaboration to identify literature for evidence synthesis. *Cell Reports Sustainability*, 1(7).

Thomas, J., McDonald, S., Noel-Storr, A., Shemilt, I., Elliott, J., Mavergames, C., and Marshall, I. J. (2021). Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for cochrane reviews. *Journal of Clinical Epidemiology*, 133:140–151.

Tran, V.-T., Gartlehner, G., Yaacoub, S., Boutron, I., Schwingshackl, L., Stadelmaier, J., Sommer, I., Aboulayeh, F., Afach, S., Meerpohl, J., et al. (2023). Sensitivity, specificity and avoidable workload of using a large language models for title and abstract screening in systematic reviews and meta-analyses. *medRxiv*, pages 2023–12.

Wang, S., Scells, H., Zhuang, S., Potthast, M., Koopman, B., and Zuccon, G. (2024). Zero-shot generative large language models for systematic review screening automation. In *European Conference on Information Retrieval*, pages 403–420. Springer.

Wolters Kluwer (2025). Ovid: Advanced search platform. Accessed: February 27, 2025.

Zhou, H., Hu, C., Yuan, Y., Cui, Y., Jin, Y., Chen, C., Wu, H., Yuan, D., Jiang, L., Wu, D., et al. (2024). Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. *IEEE Communications Surveys & Tutorials*.