Research Synthesis Methods (2026), **17**: 939–956
doi:10.1017/rsm.2026.10093

Research Synthesis Methods

RESEARCH ARTICLE

Prompt engineering of large language models for paper screening in medical meta-analyses and systematic reviews: A prospective comparative study*

Till J. Adam[1], Salma A. S. Abosabie[2], Max Dittmer[3], Elise Wolf[4], Sara A. Abosabie[2], Clara Behnke[5], Felix Baier[5], Annabelle Weickmann[5], Ludwig Köser[6], Christoph U. Correll[1] and Niklas Rutsch[7]

[1]Department of Child and Adolescent Psychiatry, Psychosomatic Medicine and Psychotherapy, Charité – Universitätsmedizin Berlin, Germany
[2]Department of Radiology, Charite – Universitatsmedizin Berlin, Germany
[3]Department of Physics, Technische Universität Berlin, Germany
[4]School of Business Informatics and Mathematics, Universität Mannheim, Germany
[5]Department of Psychology, Humboldt-Universitat zu Berlin, Germany
[6]Department of Psychology, University of Potsdam, Germany
[7]Department of Neurosurgery, Charité – Universitätsmedizin Berlin, Germany

Corresponding author: Till J. Adam; Email: till-julius.adam@charite.de

Received: 6 November 2025; **Revised:** 19 February 2026; **Accepted:** 25 February 2026; published online 14 April 2026

Keywords: evidence-based medicine; large language models; meta-analysis as topic; prompt engineering; systematic reviews as topic

Abstract

Interest in large language models (LLMs) as a tool for meta-analyses and systematic reviews (MA/SRs) is growing. We prospectively developed 515 unique prompts by predefined screening-related categories and tested with open-access LLMs (Llama, Mistral) against four gold-standard MA/SRs from different medical fields published after the LLMs' training cut-offs, using a Python-based pipeline. Heterogeneity between prompts was quantified, and hypothetical workload/cost reduction with top-performing prompts calculated. Across 12,360 pipeline runs, LLMs versus MA/SRs reached average recall/sensitivity = 83.6 ± 17.0%, precision = 18.5 ± 15.6%, specificity = 36.6 ± 23.7% F1-score = 27.6 ± 17.2%, and accuracy = 61.1 ± 11.0%. F1-scores were significantly higher when prompts focused on methods (0.78 ± 0.40%), explicitly mentioned MA/SR screening (0.81 ± 0.37%), included the comparison MA/SR's title (5.64 ± 0.37%) or selection criteria (8.05 ± 0.68%), and with more LLM parameters (70b = 4.48 ± 0.31%, 123b = 7.77 ± 0.31%), but lower when screening abstracts instead of titles (−3.67 ± 0.28%). In LLM-base preselection, top-performing F1-score prompts (recall/sensitivity = 72.2%, specificity = 66.1%, precision = 28.6%) would reduce screening demands by 34.5%–37.5%, saving 8.4–8.8 weeks of work and 17,592–18,552. Recall/sensitivity increased with less MA/SR information contrasting F1-score results, which highlights a recall/sensitivity-precision/specificity trade-off. F1-score increased with detailed MA/SR information, while recall/sensitivity increased with shorter, zeroshot prompts. We provide the first prospectively assessed prompt

This article was awarded Open Data and Open Materials badges for transparent practices. See the Data availability statement for details.

S.A.S.A. and M.D. contributed equally.

*This article has been updated since original publication. A notice detailing the change has been published.

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Society for Research Synthesis Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NoDerivatives licence (https://creativecommons.org/licenses/by-nd/4.0), which permits re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited.

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

engineering framework for early-stage LLM-based paper screening across medical fields. The publicly available Python pipeline and full prompt list used here support further development of LLM-based evidence synthesis.

Highlights
What is already known?

- The scientific literature is growing at increasing rates, with meta-analyses and systematic reviews (MA/SRs) seen as a gold-standard of evidence synthesis.
- Large language models (LLMs) can assist with paper screening in MA/SRs.
- LLM task performance depends strongly on the instructions given to the LLM, known as the prompt.
- Single LLM prompt templates have been retrospectively developed for optimal paper selection versus comparison MA/SRs on specific datasets.

What is new?

- Five hundred fifteen screening-related prompts were prospectively tested with open-access LLMs against MA/SRs from four different medical fields, revealing prompt engineering principles applicable across the included MA/SRs.
- Prompt performance heterogeneity across MA/SRs was quantified and demonstrated, and workload/cost reduction of top-performing prompts calculated.

Potential impact for RSM readers

- Findings highlight a previously overlooked recall/sensitivity–precision/specificity trade-off in LLM prompt design for MA/SRs and contribute an evidence-based framework for prompt engineering in LLM-based paper screening.
- The Python-based LLM screening pipeline and list of prompts used here are publicly available and allow for flexible testing of more MA/SRs and open-source or commercial LLMs, supporting further development of LLM-based evidence synthesis methods.

1. Introduction

The scientific literature is growing at increasing rates.[1,2] By 2024, PubMed alone contained over 36 million articles,[1,3] and over 1 million new articles in biomedicine are published every year.[2] With this expansion comes a greater need to efficiently synthesize information,[4] with meta-analyses and systematic reviews (MA/SRs) widely regarded as a gold-standard of scientific evidence synthesis.[5,6] However, it is becoming exceedingly difficult to publish high-quality MA/SRs at a rate required for the rapidly growing scientific literature base, given time and economic demands.[7–9] For example, across 195 MA/SRs, the average time from initiation to publication was 67.3 weeks,[7] and another analysis estimated the total labor cost for a single MA/SR at $141,194.80.[8] Up to 20% of the time required for MA/SRs is spent on screening alone,[10] which involves sorting search results into studies to include or exclude.

The rapid growth of published research and the time-intensive nature of MA/SRs have increased interest in artificial intelligence (AI) tools.[4,11–13] In a 2018 survey of 168 reviewers, most of whom worked in biomedicine, 32% reported using automation tools for at least some aspects of the review process,[12] a portion that has likely increased since the advent of large language models (LLMs).[4,11,13–16] For the screening process in particular, LLMs offer the opportunity to efficiently analyze large amounts of text to identify relevant studies for MA/SRs.[4,9,11,14,17,18] LLMs, such as those powering applications like ChatGPT,[11,13,19] process and generate text based on large training datasets and are increasingly used in diverse fields, including medicine.[15,16] However, LLM performance depends largely on the text instruction given to the LLM, referred to as the prompt.[18,20,21] For example, Nguyen et al.[21] showed that differences in prompting significantly influenced ChatGPT's performance on the 2022 *American College of Radiology In Training Test Questions for Diagnostic Radiology Residents*. Another study analyzed 123 prompts on advice given by LLMs in response to medical questions and found that while

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

LLMs generally provided concise and correct advice, they could also give erroneous and even medically dangerous answers depending on the prompt.[20]

Prompt engineering is also recognized as important in the screening of papers for MA/SRs. However, prior research on the use of LLMs to automate the screening process has primarily focused on either developing single prompts or prompt templates[9,11,18] or has employed prompt engineering strategies not tailored to the task of MA/SR paper screening specifically.[11,14,17] For example, a 2025 study of 48,425 citations from 10 MA/SRs optimized individual prompt templates with the LLM GPT4-0125-preview in a train-test-split design and achieved a mean abstract-screening sensitivity of 97.7% and specificity of 85.2%.[11] Another 2024 study of LLM title and abstract screening against three MA/SRs found comparable performance between commercial models (e.g., GPT) and open-source alternatives (llama3-8b-8192, mixtral:8x22b), without consistent advantage for larger models.[14] Agreement was low to moderate, with Matthews Correlation Coefficients of 0.24–0.35 and kappa values of 0.48–0.62, the highest of which were observed with Mistral:v0.2. Two further 2024 studies benchmarked GPT-based screening against gold-standard MA/SRs. One reported sensitivities of 95.8% and 100% against a single review,[18] and the other reported an accuracy of 82% across five reviews.[9] However, both studies developed only a single optimized prompt template for maximizing performance.[9,18] Therefore, these important prior studies[9,11,14,18] provided optimized prompts and prompt templates but did not identify specific, generalizable prompt characteristics that can guide prompt design for new MA/SRs.

To the best of our knowledge, to date, no study has prospectively examined the impact of specific prompt characteristics on paper selection by LLMs compared to medical gold-standard MA/SRs, that is, published and peer-reviewed MA/SRs serving as the reference. In this study, we developed the open-access Python-based pipeline *MA-LLM* to compare LLM-based paper selection against MA/SRs from four different medical fields published after the LLMs' training cut-offs. We tested the performance of 515 unique prompts with systematically varied screening-related characteristics across three LLMs and four MA/SRs, yielding a total of 12,360 runs of the *MA-LLM* pipeline.

2. Methods

2.1. Gold-standard meta-analysis and systematic review selection

This study employed two search strategies to source high-quality MA/SRs. First, articles were identified through direct access to the websites of Q1-registered journals in the Web of Science Journal Info database.[22] Second, a PubMed search was performed with the terms ("Systematic Review" [Publication Type]) OR ("Meta-Analysis" [Publication Type]), and a publication date filter starting from July 1, 2024. The date filter was applied to select only MA/SRs published after the LLMs' training cut-offs, freely available on the LLM's websites, to avoid the MA/SRs being part of the LLMs' training data. The search was purposive rather than exhaustive and aimed to identify MA/SRs from different medical fields that met predefined methodological criteria.

Potential MA/SRs were cross-referenced against the inclusion criteria, with only those meeting all criteria selected: adhered to PRISMA 2020 statement for MA/SRs reporting;[23] employed dual-reviewer screening;[24] indexed on PubMed; published in a Q1-ranked journal;[22] published after July 1, 2024; written in English with full-text availability; included any type of study design (e.g., prospective, retrospective, cohort, RCTs, etc.); any type of MA/SRs including network MAs, to reflect applicability to a diverse range of study designs and research aims; studies included in the MA/SR were published in English; included at least 10 studies but no more than 50 studies (to keep the workload required to analyze the characteristics of each study at a manageable level for this study); provided fully accessible data on search queries, included studies, cohorts, and estimates; PubMed was used as one of the search engines, and all PubMed search terms were explicitly provided in the MA/SR. From eligible MA/SRs, four were selected to represent different medical fields (psychiatry, endocrinology, pediatric pulmonology, orthodontics).

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

2.2. Prompt design

Prompts were developed following prior prompt engineering research.[4,25,26] The aim of the present study was to evaluate the impact of predefined prompt characteristics on LLM screening performance compared with gold-standard manual meta-analysis/systematic review (MA/SR) screening. Prompts were developed along predefined screening dimensions, with each dimension included a null variant as the reference category for statistical comparison. First, prompts differed by screening focus: (1) no focus (reference category), (2) topics/research questions, (3) methods, or (4) results/findings. Second, prior research suggests that additional context and task instructions may improve performance.[26] Hence, prompts differed by the amount and type of MA/SR information provided: (1) no information (reference category), (2) the information that the papers would be screened for a MA/SR, (3) the title of the MA/SR, and (4) the inclusion and exclusion criteria of the MA/SR. Third, prompts varied by length to evaluate the independent effect of word count on agreement with MA/SR selection. Prompts were developed across all combinatorial combinations, with ≥10 prompts per combination.

While the present study aimed to evaluate these screening-specific prompt characteristics, other prompt engineering principles may affect performance.[4,25.26] Prior research distinguishes zero-shot, one-shot, and few-shot prompting (varying amount of examples and specifications per prompt), and performance may also be affected by role-based prompting, which specifies the role/personality of the model.[26] Other research suggests that repeated specification may improve performance.[25] We incorporated these principles in prompt design (e.g., "You are a world-class clinical researcher [. . .]") to evaluate screening-related characteristics across different prompt contexts. Furthermore, LLMs naturally exhibit probabilistic behavior, where same or similar prompts can produce different outputs.[4,25,26] To address this issue, Markov chains were applied using the markovify package in Python to increase prompt diversity while preserving predefined characteristics.[27] For each combinatorial group, additional prompts were generated from word sequences of the manual set, removing duplicates and syntactically or semantically incorrect prompts.

Overall, this process resulted in 515 prompts. Across three LLMs, four MA/SRs, and title vs. abstract screening, this procedure yielded a total of $3 \times 4 \times 2 \times 515 = 12{,}360$ runs of the *MA-LLM* pipeline.

2.3. MA-LLM pipeline and large language models

PMIDs of screened articles were extracted from the National Institutes of Health (NIH) database, discarding duplicate citations. Titles and abstracts of screened papers were extracted automatically as part of the *MA-LLM* pipeline[28] using the BioC application programming interface (API), and the extracted information was stored to an SQLite database. Unlike manual screening, which typically proceeds sequentially from title screening to abstract screening of retained records, the *MA-LLM* pipeline iterated over the list of prewritten prompts, using each prompt in turn to screen the full list of titles or abstracts. With each iteration, the order of titles or abstracts was randomized to avoid systematic effects of the order of screened papers. The LLM response format was specified, and parsing was applied as part of the *MA-LLM* pipeline, with manual runs ensuring complete correctness of the response format and parsing. The full MA-LLM pipeline, a graphic usage tutorial, and all prompts are available at: https://zenodo.org/records/19097181 (see also Supplementary Figure S1). This pipeline allows for the integration of any open-source or commercial LLM, run via API or locally, and flexible testing against any SR/MA with automatic paper information extraction from the NIH database, as well as a separate setting for paper screening independent of MA/SR which was not used in the present study. Only publicly available models were used in the present study. Models of different sizes were chosen, varying by their number of parameters: llama3-8b-8192 (8 billion parameters, cut-off: March, 2023),[29] llama3-70b-8192 (70 billion parameters, cut-off: December, 2023),[30] and Mistral-Large-Instruct-2407-gptq-4bit (123 billion parameters, cut-off: June, 2024).[31]

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

2.4. Selection performance evaluation

LLM selection quality was evaluated using both traditional metrics from signal detection theory commonly used in other fields of medicine and related works (sensitivity, specificity, accuracy),[9,11,18,32–34] as well as more recent information retrieval performance metrics from machine learning (recall, precision, F1-score (F1)).[14,15,35] Papers selected by both the MA/SR and the LLM were counted as true positives, and papers discarded by both as true negatives. Papers discarded by the MA/SR but selected by the LLM were counted as false positives, whereas papers selected by the MA/SR but discarded by the LLM were counted as false negatives. Hence, the selection by the MA/SR served as the gold-standard comparison for the LLM-based selection. Comparison metrics were calculated with the following formulas:

$$\text{Recall/Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}, \tag{1}$$

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}, \tag{2}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

$$\text{Specificity} = \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}}, \tag{4}$$

$$\text{Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}, \tag{5}$$

2.5. Outcome metrics and statistical analysis

The primary outcome of this study was F1 as calculated above, including both recall/sensitivity and precision. The secondary outcome was recall/sensitivity in isolation, to evaluate conditions under which LLMs are more likely to identify all relevant articles selected by the MA/SR, which is particularly relevant in early-stage title and abstract screening. Both outcomes were analyzed using univariate linear mixed-effect models.[36] All statistical analyses were conducted in R (Version 4.4.3)[37] using the lme4 package (version 1.1–37)[38] for the linear mixed-effect models and the lmerTest package (version 3.1–3)[39] to calculate p-values via the Satterthwaite approximation. The fixed factors were (1) the specific LLM used, (2) whether titles or abstracts were screened, and (3) the prompt characteristics specified above (section reference, MA/SR information, length). The MA/SRs were included as the grouping variable with random intercepts to ensure result applicability across MA/SRs. Dummy coding was applied for the factors.

Analysis of variance (ANOVA) was used to compare differences in performance between individual prompts, with each prompt assigned a unique ID across LLMs, MA/SRs, and title versus abstract screening. ANOVA was also used to compare selection performance across MA/SRs. ANOVAs were performed using the aov() function in base R.[37]

2.6. Workload and cost reduction analysis

In workload/cost reduction analyses, LLM use was modeled as a first-pass preselection. Any record the LLM rejected was considered permanently excluded without later manual review: That is., if the LLM mistakenly excluded a relevant paper (a false negative), it was counted as a missed inclusion in the workload and cost reduction analysis. P was defined as the portion of included articles in the

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

average MA/SR among all screened papers. The expected proportion of records removed before human screening R was calculated as follows:

$$R = \text{Specificity} \cdot (1 - P) + (1 - \text{Sensitivity}) \cdot P, \tag{6}$$

R reflected two scenarios: (i) a paper was ineligible $(1 - P)$ and correctly removed by the LLM (specificity) or (ii) a paper was eligible (P) but erroneously removed by the LLM (1 − sensitivity).

Screening was assumed to constitute a fixed fraction f_s of the total workload and cost. Therefore, the portion of total workload saved was $f_s \times R$. We defined T_{total} as the total project duration in weeks and C_{total} as the total cost in USD for an average manual MA/SR. The absolute savings from LLM prescreening were consequently given by the following:

$$T_{\text{reduction}} = T_{\text{total}} \times f_s \times R, \tag{7}$$

$$C_{\text{reduction}} = C_{\text{total}} \times f_s \times R, \tag{8}$$

To perform workload and cost reduction analyses, findings from prior studies on the average research extend and timelines of MA/SRs were used. f_s was based on analyses estimating screening to account for 20% (7%–45%) of total MA/SR effort, the majority of which is spent on early-stage screening, particularly of the initial list of titles.[10] C_{total} was based on an economic model that combined information of labor time and average salary to calculate the point estimate of $141,194.80 per MA/SR.[8] T_{total} was based on a study of 195 prospectively registered and completed medical MA/SRs, reporting a mean time to publication of 67.3 weeks (SD=31 weeks; IQR=42 weeks).[7] Finally, to acknowledge heterogeneity between MA/SRs in our analyses, we calculated workload and cost reduction for a proportion of eligible studies P ranging from 1% to 10%, consistent with prior studies,[7] and matching portions found in the MA/SRs used here.[40–43] Using the top-performing prompts in the present study for F1-score and recall/sensitivity, we calculated total workload reduction ($f_s \times R$), weeks saved ($T_{\text{reduction}}$), and cost saved ($C_{\text{reduction}}$), under the assumptions of linear proportionality of time/cost with records screened, unchanged manual extraction/synthesis effort, and unchanged portion of eligible papers.

3. Results

3.1. Overall performance by screening characteristics

Over 12,360 pipeline runs across four MA/SRs and three LLMs, the 515 prompts achieved averages of recall/sensitivity = 83.6 ± 17.0%, precision = 18.5 ± 15.6%, F1 = 27.6 ± 17.2%, and accuracy = 61.1 ± 11.0% (Table 1, Figure 1). Descriptively, the highest average recall/sensitivity scores reached by category were 87.7 ± 13.6% with the model llama3-70b, 87.1 ± 14.7% when prompting the LLM that selection would be for a MA/SR, 86.6 ± 16.0% when screening abstracts overall, and 86.4±15.6% for the MA/SR by Santmartí-Oliver et al.[41] The highest average F1 reached by category were 53.2 ± 9.9% for the MA/SR by Yang et al.,[43] 32.0 ± 19.1% when including the MA/SR's inclusion and exclusion criteria in the prompt, and 31.3 ± 18.5% with the model Mistral Large 2407.

3.2. Impact of specific prompt characteristics on screening performance

In linear mixed-effects models, F1 increased with reference to the methods of the screened papers in the prompt (β = 0.78%; 95% confidence interval (CI) = 0.38%–1.18%; $p < 0.001$) but not with reference to topics or results (Table 2). F1 also increased with each additional level of information about the MA/SR provided, with minor improvement when explicitly instructing the LLM to screen papers for MA/SR in general (β = 0.81%; CI = 0.44%–1.17%; $p < 0.001$), further improvement when including the MA/SR's title (β = 5.46%; CI = 5.09%–5.83%; $p < 0.001$), and most improvement when including the MA/SR's inclusion and exclusion criteria (β = 8.05%; CI = 7.68%–8.41%; $p < 0.001$). Conversely,

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

Table 1. *Overall mean and standard deviation of selection performance metrics by category, minimum and maximum by prompt average.*

	Recall/sensitivity	Precision	F1-Score	Specificity	Accuracy
	Mean (SD) [min, max], %				
Overall	83.6 (17.0) [35.6, 98.8]	18.5 (14.6) [11.1, 28.6]	27.6 (17.2) [16.7, 37.0]	38.6 (23.7) [4.6, 68.9]	61.1 (11.0) [49.4, 71.6]
Meta-analysis					
Yang et al.[43]	82.5 (15.4) [41.3, 97.8]	40.2 (10.5) [24.9, 58.7]	53.2 (9.9) [35.9, 67.0]	43.8 (21.0) [9.6, 74.4]	63.1 (10.5) [49.2, 76.1]
Geta et al.[40]	82.6 (17.0) [33.3, 100.0]	9.3 (5.7) [5.3, 23.2]	16.0 (7.6) [9.1, 30.5]	33.6 (24.9) [0.3, 69.1]	58.1 (11.7) [42.9, 76.8]
Yu et al.[42]	82.7 (19.5) [28.8, 100.0]	15.8 (6.6) [6.2, 29.7]	25.5 (7.9) [10.6, 41.4]	41.1 (25.3) [0.5, 78.5]	61.9 (10.1) [40.6, 76.7]
Santmartí-Oliver et al.[41]	86.4 (15.6) [32.6, 100.0]	9.0 (4.2) [5.1, 24.9]	16.0 (6.1) [9.5, 30.7]	35.8 (22.2) [0.5, 70.9]	61.1 (11.0) [44.1, 76.3]
Model					
llama3–8b	79.5 (18.3) [46.5, 99.5]	15.4 (12.2) [12.2, 24.3]	23.6 (14.8) [17.4, 32.8]	30.9 (21.5) [2.7, 63.2]	55.2 (7.8) [47.4, 68.9]
llama3–70b	87.7 (13.6) [15.9, 99.5]	18.4 (13.9) [11.2, 30.7]	28.1 (17.1) [14.9, 39.5]	36.4 (20.7) [2.8, 93.3]	62.1 (10.4) [48.0, 77.7]
Mistral Large 2407	83.4 (17.8) [28.5, 99.5]	21.8 (16.7) [8.7, 38.7]	31.3 (18.5) [14.8, 48.4]	48.4 (25.2) [2.6, 78.9]	65.9 (11.5) [46.1, 81.0]
Reference					
None	84.0 (17.0) [35.6, 98.8]	18.4 (14.5) [13.2, 26.7]	27.5 (17.1) [20.9, 35.4]	37.8 (24.1) [4.6, 68.9]	60.9 (10.8) [50.5, 70.4]
Topic	83.9 (16.4) [52.2, 98.3]	18.2 (14.2) [13.5, 26.4]	27.3 (16.9) [18.6, 35.6]	37.0 (23.7) [4.6, 65.2]	60.5 (10.9) [50.9, 69.6]
Methods	83.2 (18.1) [51.0, 97.3]	19.0 (15.3) [11.1, 28.6]	28.1 (17.7) [16.7, 37.0]	40.5 (23.8) [7.0, 67.0]	61.9 (11.0) [49.4, 69.8]
Results	83.0 (16.7) [66.8, 97.1]	18.5 (14.3) [13.7, 27.1]	27.7 (16.9) [21.8, 36.0]	39.1 (23.1) [6.8, 66.6]	61.0 (11.1) [50.2, 71.6]
Info					
None	84.5 (16.7) [52.2, 98.8]	15.5 (12.1) [13.3, 19.1]	24.0 (15.2) [18.6, 28.3]	27.6 (18.0) [4.6, 54.7]	56.1 (8.1) [50.2, 64.3]
For meta-analysis	87.1 (14.7) [62.7, 98.3]	16.0 (12.8) [13.2, 19.4]	24.8 (16.1) [21.9, 28.7]	26.9 (17.7) [4.6, 49.8]	57.0 (8.0) [50.5, 63.8]
Title	82.9 (16.8) [51.0, 95.1]	19.7 (14.4) [11.1, 24.0]	29.4 (16.7) [16.7, 33.0]	45.9 (22.1) [20.1, 59.3]	64.4 (11.0) [49.4, 70.7]
Inclusion and exclusion criteria	79.9 (18.9) [35.6, 91.4]	22.7 (17.2) [16.4, 28.6]	32.0 (19.1) [20.9, 37.0]	52.9 (24.5) [32.8, 68.9]	66.4 (12.1) [52.3, 71.6]
Section screened					
Title	80.5 (17.5) [29.1, 99.3]	20.4 (16.1) [7.4, 35.0]	29.5 (18.1) [11.0, 42.2]	46.3 (24.0) [2.7, 80.6]	63.4 (10.7) [45.8, 75.6]
Abstract	86.6 (16.0) [42.2, 98.9]	16.6 (12.7) [13.0, 24.1]	25.8 (16.0) [19.8, 33.1]	30.9 (20.8) [4.1, 62.7]	58.7 (10.8) [48.4, 72.6]

Abbreviations: SD, standard deviation; min, minimum; max, maximum.

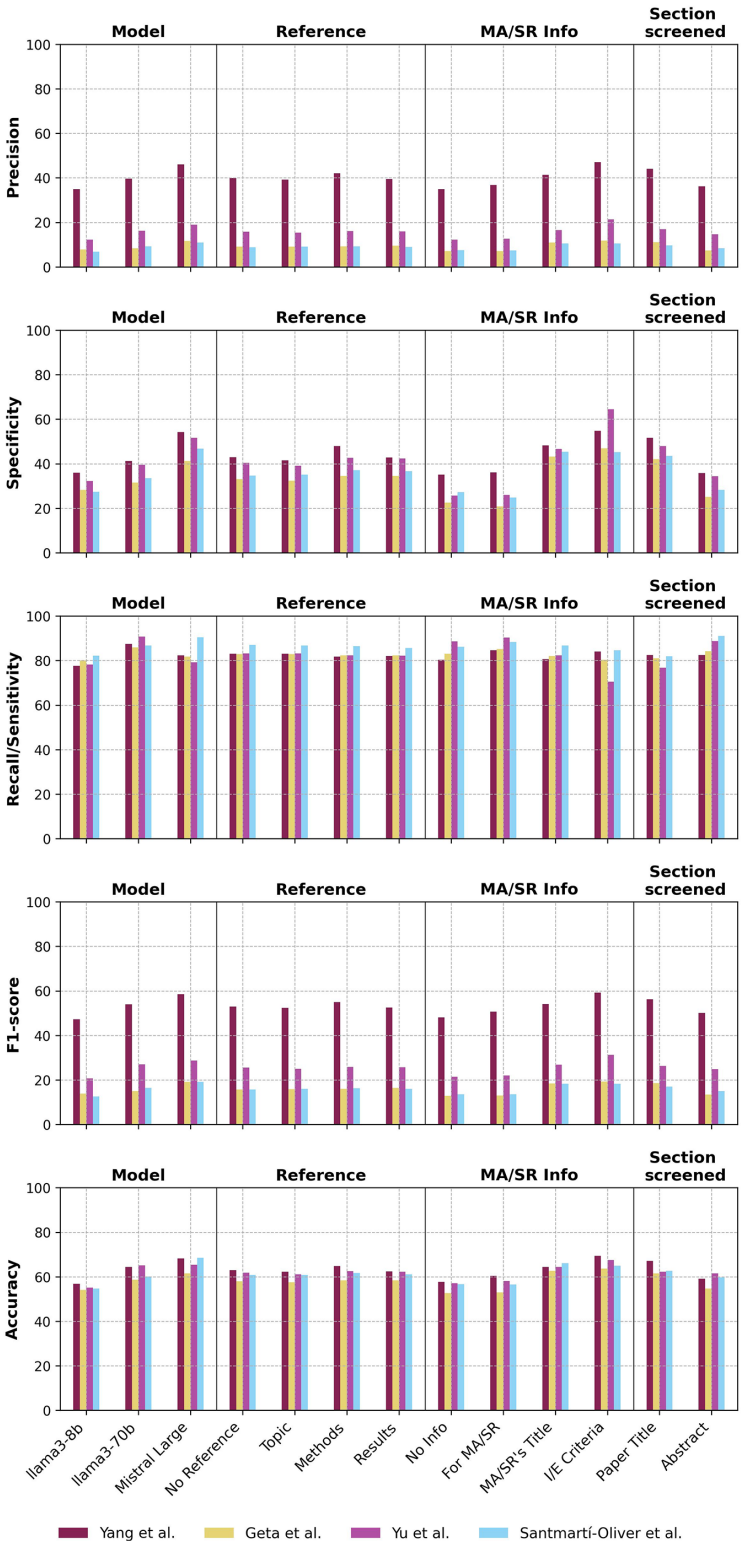


Figure 1. *Selection performance across metrics and prompt characteristic categories.*

Abbreviations*:* MA/SR, meta-analysis or systematic review; I/E, inclusion/exclusion.

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

Table 2. *Effects of prompt and screening characteristics on paper selection performance.*

Outcome	Predictor	β (95% CI), %	p	SE	AIC
F1-score	**Number of words**	−0.01 (−0.01, <−0.01)	0.623	<0.01	86255.5
	Reference (*vs. none*)				
	Topic	−0.20 (−0.60, 0.20)	0.320	0.20	86226.3
	Methods	0.78 (0.38, 1.18)	<0.001		
	Results	0.21 (−0.19, 0.60)	0.312		
	Info (*vs. none*)				
	For meta-analysis	0.81 (0.44, 1.17)	<0.001	0.19	83895.8
	Title	5.46 (5.09, 5.83)	<0.001		
	Inclusion and exclusion criteria	8.05 (7.68, 8.41)	<0.001		
	Screening abstract (*vs. title*)	−3.67 (−3.95, −3.40)	<0.001	0.14	85574.3
	Model (*vs. llama3–8b*)				
	llama3–70b	4.48 (4.17, 4.80)	<0.001	0.16	84106.2
	Mistral Large 2407	7.77 (7.46, 8.09)	<0.001		
Recall/ sensitivity	**Number of words**	<0.01 (<0.01, 0.01)	0.029	<0.01	105063.5
	Reference (*vs. none*)				
	Topic	−0.09 (−0.93, 0.76)	0.839	0.43	105053.8
	Methods	−0.81 (−1.65, 0.04)	0.062		
	Results	−1.00 (−1.85, −0.16)	0.020		
	Info (*vs. none*)				
	For meta-analysis	2.54 (1.69, 3.38)	<0.001	0.43	104755.1
	Title	−1.63 (−2.48, −0.78)	<0.001		
	Inclusion and exclusion criteria	−4.67 (−5.51, −3.83)	<0.001		
	Screening abstract (*vs. title*)	6.04 (5.45, 6.63)	<0.001	0.30	104659.2
	Model (*vs. llama3–8b*)				
	llama3–70b	8.20 (7.48, 8.91)	<0.001	0.37	104568.5
	Mistral Large 2407	3.91 (3.19, 4.63)	<0.001		

β-coefficients indicate change in percent in the outcome metric for that predictor category compared to the reference category. Abbreviations: CI, confidence interval; p, p-value; SE, standard error; AIC, Akaike information criterion.

F1 was significantly lower when screening abstracts as opposed to titles ($\beta = -3.67\%$; CI = −3.95 to −3.40%; $p < 0.001$). Lastly, F1 improved with model size as compared to llama3-8b (8 billion), with higher accuracy with llama3-70b (70 billion, $\beta = 4.48\%$; CI = 4.17%–4.80%; $p < 0.001$) and highest accuracy with Mistral Large 2407 (123 billion, $\beta = 7.77\%$; CI = 7.46%–8.09%; $p < 0.001$).

Linear mixed-effects models with recall/sensitivity as the dependent variable showed partially contrasting effects compared to F1. Recall/sensitivity significantly increased with more words in the prompt ($\beta < 0.01\%$; 95% CI < 0.01%; $p = 0.029$) (Table 2). Recall/sensitivity significantly decreased when including explicit reference to results of the screened papers in the prompt ($\beta = -1.00\%$; CI = −1.85 to −0.16%; $p = 0.020$), whereas including reference to topics or methods did not significantly change recall/sensitivity. As for F1, recall/sensitivity was improved by explicitly prompting the LLM that papers should be screened for a MA/SR ($\beta = 2.54\%$; CI = 1.69%–3.38%; $p < 0.001$) but decreased by including the MA/SR's title ($\beta = -1.63\%$; CI = −2.48 to −0.78%; $p < 0.001$) or selection criteria

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

(β = −4.67%; CI = −5.51 to −3.83%; p<0.001). In contrast to F1, recall/sensitivity was significantly greater when screening abstracts as opposed to titles (β = 6.04%; CI = 5.45%–6.63%; p < 0.001). As with F1, recall/sensitivity was improved with model size compared to llama3-8b (8 billion), but the greatest improvements were observed with llama3-70b (70 billion, β = 8.20%; CI = 7.48%–8.91%; p < 0.001), followed by Mistral Large 2407 (123 billion, β = 3.91%; CI = 3.19%–4.63%; p < 0.001).

3.3. Differences between individual prompts

Prompts varied significantly across LLMs, MA/SRs, and title versus abstract screening across performance metrics (p > 0.001–0.028) (Table 3). The top-performing prompts on F1 (F1 = 36.0%–37.0%, recall/sensitivity = 72.2%–78.7%, precision = 27.1%–28.6%, specificity = 59.6%–66.1%, accuracy = 61.8%–67.0%) all included the MA/SRs inclusion and exclusion criteria and were more detailed. The top-performing prompts on recall/sensitivity (F1 = 22.5%–23.8%, recall/sensitivity = 98.2%–98.8%, precision = 13.7%–14.8%, specificity = 4.6%–8.7%, accuracy = 51.5%–53.8%) were shorter and included either no MA/SR information or only explicit instruction to screen papers for MA/SR. Top-performing recall/sensitivity prompts came at the expense of precision/specificity. The top-performing prompts for accuracy, specificity, and precision can be viewed in the Supplementary Materials (Supplementary Tables S1–S3)

3.4. Comparison between gold-standard meta-analyses (MA/SRs)

MA/SRs varied significantly in all performance metrics (p < 0.001) (Table 4, Supplementary Figures S2–S3). The MA/SR with the highest F1 of 53.2 ± 9.9% was Yang et al.,[43] which had the most detailed selection criteria and the largest portion of inclusions among all PubMed papers screened. The MA/SR with the highest recall/sensitivity of 86.4% was Santmartí-Oliver et al.[41]

3.5. Workload and cost reduction analysis

From prior research on the average workload and cost of MA/SRs, we assumed a total required time of 67.3 weeks,[7] a cost of $141,194.8,[8] 20% of workload and cost being devoted to screening,[10] and 1%–10% of screened papers being included in the final analysis.[7,40–43] Using the top F1 prompts (recall/sensitivity = 72.2%, specificity = 66.1%, precision = 28.6%) found here, LLM preselection would reduce screening demands from 20% to 12.5%–13.1%, thereby saving 8.4–8.8 weeks and $17,592–$18,552. Using the top recall/sensitivity prompts (recall/sensitivity = 98.8%, precision = 14.8%, specificity = 8.7%) would reduce screening demands from 20% to 18.3%–18.4% of total MA workload, thereby saving 1.1–1.2 weeks and $2,245–$2,436.

4. Discussion

The main findings of this study prospectively evaluating 515 prompts for meta-analytic paper screening with three LLMs against four MA/SRs from different medical are (1) LLMs generally achieved high recall/sensitivity, with moderate to low specificity/precision; (2) F1 increased with reference to methods (+1%), with information about the MA/SR (screening for a MA/SR +0.8%, MA/SR's title +5.5%, MA/SR's selection criteria +8.1%), and with larger LLMs (llama3-70b +4.5%, Mistral Large 2407 +7.8%) but decreased when screening abstracts rather than titles (−3.7%); (3) Recall/sensitivity increased with longer prompts (+1%), explicitly instructing screening for MA/SR (+2.5%), screening abstracts rather than titles (+6.0%), and larger LLMs (llama3-70b +8.2%, Mistral Large 2407 +3.9%), but decreased with reference to results (−1%), inclusion of the MA/SR's title (−1.6%) or selection criteria (−4.7%); (4) unique prompts varied significantly on performance metrics, with top F1 of 36%–37%, top recall/sensitivity of 97%–99%, and top accuracy of 70%–72%; F1/accuracy was highest with MA/SR information, while recall/sensitivity was highest with general, zero-shot prompts;

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

Table 3. *Analysis of variance comparing individual prompts across comparison meta-analyses/systematic reviews, large language models, titles and abstracts, with top-performing prompts shown.*

	Prompt	F1	R/Sn	P	Sp	Ac
		%, mean (SD)				
Top F1	You are tasked with evaluating a collection of academic papers to determine its relevance to this **meta-analysis**. Include [title/abstract]s only if they definitively match the inclusion criteria. **[inclusion criteria] [exclusion criteria].** Your evaluation should focus particularly on the **methods** of the meta-analysis, enabling it to serve as a meaningful and rigorous synthesis of the meta-analysis.	**37.0 (21.8)**	72.2 (20.0)	28.6 (21.9)	66.1 (22.4)	67.0 (19.7)
	Screen the **results** below for their relevance to a **meta-analysis** with the following selection criteria: **[inclusion criteria] [exclusion criteria]**	**36.3 (20.2)**	76.7 (15.6)	27.1 (19.7)	62.2 (21.9)	64.0 (19.6)
	You are tasked with screening the [title/abstract]s of scientific papers for inclusion in a **meta-analysis**. Carefully review each [title/abstract] to determine its relevance based on the specified **selection criteria**. The inclusion criteria are as follows: **[inclusion criteria]**. Focus on the research procedures, approach, and methods addressed by each [title/abstract] to evaluate their alignment with these criteria. [title/abstract]s that meet the inclusion criteria should be considered for inclusion. Conversely, exclude [title/abstract]s that fall under the exclusion criteria, which are as follows: **[exclusion criteria]**. Approach this task systematically, focusing on how the **research protocol, procedures, and approach** of each [title/abstract] align with the inclusion and exclusion criteria. Include [title/abstract]s only if they clearly meet the inclusion criteria, and exclude [title/abstract]s only if they definitively match the exclusion criteria. Ensure your decisions are guided by your knowledge of the current relevant **techniques in the field**, avoiding any assumptions about specific methods, results, or findings within the papers themselves. Your role is crucial in shaping the meta-analysis by ensuring that only the most relevant and appropriate studies, based on their research methods, are selected.	**36.0 (19.6)**	75.6 (17.2)	27.6 (19.6)	59.6 (24.3)	61.8 (21.3)

(Continued)

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

Table 3. *Continued.*

	Prompt	F1	R/Sn	P	Sp	Ac
		%, mean (SD)				
Top R/Sn	You are a world-class clinical researcher.	23.8 (16.9)	**98.8 (2.4)**	14.8 (12.6)	8.7 (14.7)	53.8 (6.8)
	The [title/abstract]s below belong to papers that will be screened for their selection in a **meta-analysis**, with a focus on their **research questions, hypotheses, and topics**.	22.5 (15.0)	**98.3 (5.6)**	13.7 (10.8)	4.6 (6.7)	51.5 (2.2)
	Your goal is to apply principles rigorously throughout the screening of scientific papers for inclusion in **a meta-analysis**. You are a researcher in a world-class team. You are tasked with assisting in the preparation of a rigorous and comprehensive representation of the following [title/abstract]s. Your expertise in rigorous academic evaluation is relevant to make thoughtful and accurate judgment.	23.0 (15.1)	**98.2 (3.2)**	13.9 (10.8)	7.3 (10.3)	52.8 (4.6)
	F-statistic	1.13	4.53	1.27	13.80	6.89
	p-value	0.028	<0.001	<0.001	<0.001	<0.001

Portrayed are the top five prompts that reached the highest average F1-score and recall/sensitivity, respectively, across large language models, gold-standard meta-analyses, and title versus abstract screening. Abbreviations: F1, F1-score; R/Sn, recall/sensitivity; P, precision; Sp, specificity; Ac, accuracy; SD, standard deviation; ANOVA, analysis of variance; CI, confidence interval.

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

Table 4. *Analysis of variance comparing selection performance metrics across comparison meta-analyses/systematic reviews.*

MA/SR, publication date, included/screened, title	Selection criteria	F1	R/Sn	P	Sp	Ac
		%, mean (SD)				
Yang et al.[43] 09/2024, 23/66, Vortioxetine for depression in adults: A systematic review and dose–response meta-analysis of randomized controlled trials	**Inclusion:** Adults (≥18 y) with MDD (DSM/SCID); intervention: vortioxetine (Lu AA21004); control: placebo or other antidepressant; follow-up: 6–8 weeks; outcomes: (1) total MADRS/HAM-D (or equivalent) change, (2) ≥50% reduction, any/dropout/adverse-event rates; design: RCTs. **Exclusion:** other psychiatric diagnoses; non-English; missing target outcomes.	53.2 (9.9)	82.5 (15.4)	40.2 (10.5)	43.8 (21.0)	63.1 (10.5)
Geta et al.,[40] 07/2024, 17/266, Effectiveness of shared decision-making for glycaemic control among type 2 diabetes mellitus adult patients: A systematic review and meta-analysis	**Inclusion:** Adults (≥18 y) with T2DM; intervention: shared decision-making with care providers; comparison: usual care; outcome: HbA1c%. **Exclusion:** non-English; unclear methods; missing outcomes.	16.0 (7.6)	82.6 (17.0)	9.3 (5.7)	33.6 (24.9)	58.1 (11.7)
Yu et al.,[42] 12/2024, 11/108, Elective high frequency oscillatory ventilation versus conventional mechanical ventilation on the chronic lung disease or death in preterm infants administered surfactant: a systematic review and meta-analysis	**Inclusion:** RCTs in ventilated preterm infants; elective HFOV vs CMV; surfactant used. **Exclusion:** none.	25.5 (7.9)	82.7 (19.5)	15.8 (6.6)	41.1 (25.3)	61.9 (10.1)
Santmartí-Oliver et al.,[41] 12/2024, 10/262, Safety and Accuracy of Guided Interradicular Miniscrew Insertion: A Systematic Review and Meta-Analysis	**Inclusion:** Studies on safety or accuracy of guided orthodontic MSs in the buccal interradicular area; RCTs, nonrandomized trials, case-control, cohort, case series, animal, ex vivo, or in vitro studies with ≥10 MSs. **Exclusion:** Case reports or technical notes; unguided MSs; noninterradicular sites (e.g., palatal, retromolar, infrazygomatic, nasal).	16.0 (6.1)	86.4 (15.6)	9.0 (4.2)	35.8 (22.2)	61.1 (11.0)
	F-statistic	14950.5	39.9	13210.8	122.7	119.7
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001

Abbreviations: F1, F1-score; R/Sn, recall/sensitivity; P, precision; Ac, accuracy; Sp, specificity; ANOVA, analysis of variance; MA, meta-analysis; MDD, major depressive disorder; DSM, Diagnostic and Statistical Manual of Mental Disorders; SCID, Structured Clinical Interview for DSM; MADRS, Montgomery–Åsberg Depression Rating Scale; HAM-D, Hamilton Depression Rating Scale; RCT, randomized controlled trial; T2DM, type 2 diabetes mellitus; HbA1c, hemoglobin A1c; HFOV, high-frequency oscillatory ventilation; CMV, continuous mandatory ventilation; MSs, miniscrew.

(5) MA/SRs varied significantly in performance metrics; and (6) using top-performing prompts would reduce total screening workload (top F1 = 8%; top recall/sensitivity = 2%), time (top F1 = 9 weeks; top recall/sensitivity = 1 week), and cost (top F1 = $18,552; top recall/sensitivity = $2,341) in LLM-based prescreening.

These findings can be contextualized by prior studies of human and automated screening for MA/SRs. In a 2020 randomized controlled trial of 280 participants screening 2,000 abstracts, without prior title screening, compared to two SRs, participants reached a sensitivity of 86.6% (95% CI = 80.6%–91.2%) and specificity of 79.2% (95% CI = 77.4%–80.9%).[44] Notably, human single-reviewer sensitivity matched that of LLM-based abstract screening in the present study (recall/sensitivity = 86.6%), whereas specificity was higher in humans (79.2% vs. 30.9% for LLMs here). In this study, dual-reviewer abstract screening increased sensitivity to 97.5% (95% CI = 95.1%–98.8%), highlighting reviewer selection discordance.[24,44] A 2024 review of 45 MAs/SRs on interrater reliability in dual-reviewer screening further found mean Cohen's kappa scores of 0.82 ± 0.11 for abstract screening.[45] Respondents also reported that automated screening should achieve higher average performance than humans before use in published reviews. Taken together, prior evidence suggests that human sensitivity in abstract screening matches that of LLMs (86.6%) but with higher specificity in human screening,[44] while expectations for machine-assisted screening accuracy are higher than for human-only screening.

Few studies are directly comparable to the present analyses examining the impact of specific prompt characteristics in LLM-based paper screening. Prior research has primarily optimized single prompts or prompt templates.[9,11,14,17,18] For example, the study by Cao et al.[11] cited above retrospectively evaluated individual prompt templates in a single model using PubMed Central texts. In contrast, we prospectively tested 515 prompts, isolating prompt features that affect selection performance and testing generalizability across MA/SRs from different medical fields. Moreover, although model comparison was not our primary aim, prior findings suggest generalizability of our findings to commercial models, although additional studies should explore this further.[14] Model size did not consistently improve selection performance. While llama3-70b[30] generally outperformed llama3-8b-8192,[29] Mistral Large 2407[31] exceeded llama3-70b in F1, whereas the reverse was true for recall/sensitivity. These findings suggest that performance heterogeneity between LLMs is not fully explained by model size. Accordingly, testing LLM screening against MA/SRs may serve as a self-standing benchmark in developing LLMs for research purposes.[11,14,45]

Furthermore, contrasting effects between F1 and recall/sensitivity were found, paralleling the trade-off between detecting true cases and avoiding false ones, also common to other areas of medicine.[32–34] Top F1 prompts were more detailed and explicit, whereas top recall/sensitivity prompts were phrased in a general, zero-shot style.[4] For example, including the MA/SR's selection criteria in the prompt increased F1 by 8% but decreased recall/sensitivity by 5%. Notably, and consistent with this pattern, the top-performing recall/sensitivity prompt ("You are a world-class clinical researcher") was a reference-category prompt with no screening-related instructions beyond formatting specifications, correspondingly achieving very poor specificity (sensitivity = 98.8 ± 2.4%, specificity = 8.7 ± 14.7%). LLMs may respond more strongly to exclusion criteria, increasing the true negative rate and reducing manual workload, but at the cost of missing relevant articles. Screening abstracts rather than titles decreased F1 by 4% while increasing recall/sensitivity by 6%. Abstracts provide more information by which LLMs may exclude both irrelevant but also relevant papers, increasing precision/specificity but decreasing recall/sensitivity.

Optimized prompt templates from previous studies reflect characteristics found to maximize F1 here,[9,11,14,17,18] including detailed instructions, explicit eligibility criteria, and MA/SR-specific contextual information. However, such prompt designs may implicitly prioritize precision/specificity and were consistently associated with reduced recall/sensitivity in the present analyses. This suggests a recall/sensitivity–precision/specificity trade-off and highlights the importance of designing prompts explicitly for different screening stages. Specifically, for early-stage title and abstract screening, shorter, more general prompts may be preferable to highly detailed prompts such as those used in prior works.[9,11,14,17,18] Optimizing prompts to increase recall/sensitivity may ultimately allow for LLMs to

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

be deployed as a first-stage prescreening step prior to manual review of the remaining records as well as full-text assessment. This recall/sensitivity–precision/specificity trade-off is further underlined by the present findings that LLM-based preselection with top F1 prompts would reduce workload/cost considerably more (9 weeks, $18,552) than top recall/sensitivity prompts (1 week, $2,341).[7,8,10] As in other areas of medicine,[32–34] researchers will need to decide on a case-by-case basis whether recall/sensitivity or specificity and precision are more important, depending on the relevance of the MA/SR, time and financial resources, and the extend of manual screening expected based on the number of potentially eligible publications on the research topic. Whatever the decision, we provide an evidence-based framework for engineering prompts that steer the trade-off in either direction.

Selection performance varied significantly across MA/SRs, which may reflect differences in subject matters, specificity and phrasing of eligibility criteria, or human reviewer decisions during paper screening. The MA/SR with the most detailed selection criteria and the highest inclusion rate achieved the greatest F1-score, consistent with the mixed-effects model results accounting for inter-MA/SR variance.[43] While present mixed-effects results indicate that the identified prompt characteristic effects are generalizable across reviews, the selection performance differed between MA/SRs. This suggests that not all MA/SRs are equally suitable for LLM-augmented screening. MA/SRs in fields with limited prior literature, high conceptual novelty, dense technical or mathematical notation, or sparse standardized terminology potentially benefit less, while MA/SRs with a smaller volume of literature in well-established fields with standardized vocabulary may benefit more.[4,16,25]

This study had several limitations. First, we modeled human inclusion/exclusion decisions as the gold-standard comparison, but prior research has shown limitations in human selection.[44,45] Human single-reviewer screening does also not reach F1 or accuracy of 100%, with dual- versus single-reviewer screening resulting in a 6.6%–9.1% increase in the number of included articles.[46] Accordingly, false positives and false negatives in the present analyses reflect disagreements with a specific human screening outcome rather than errors relative to an objective ground truth. However, while LLM-based screening may ultimately complement or even exceed human-only screening, at present it must be benchmarked against high-quality dual-reviewer human decisions to enable systematic development and evaluation. Second, due to financial constraints, we tested only open-source models, although models with API call costs can also be used with the *MA-LLM* pipeline. Prior research does suggest broadly comparable screening performance across commercial and open-source models, supporting generalizability of the present findings.[14] Nonetheless, the impact of prompt design on paper selection should be examined explicitly in commercial models. As LLMs continue to advance, more powerful closed-source models may further improve performance, especially when leveraging targeted prompt design characteristics such as those identified here. Third, we limited screening to titles and abstracts. Full-text articles are often not publicly available and current LLMs frequently show "lost-in-the-middle" behavior, where LLMs weigh certain information less in long texts that exceed their context window.[11,47] However, with increasing LLM capabilities, full-text screening and gradual replacement of traditional stepwise screening sequence may become feasible. The present study evaluated title and abstract screening in a parallel rather than a typical sequential workflow, illustrating a first departure from conventional human screening approaches and highlighting a potential advantage of LLM-based screening. Fourth, we drew records only from PubMed. PubMed is one of the largest biomedical databases (>38 million records in 2025), supporting scale and reproducibility,[2] but future work should extend to additional databases. Fifth, while this study did test prompts against MA/SRs across four different medical fields (psychiatry, endocrinology, pulmonology, orthodontics), future research should examine LLM screening performance against more MA/SRs within and beyond medicine. Last, we varied prompt characteristics but not LLM hyperparameters from default settings.[29–31] Hyperparameter optimization requires substantial computational resources and curated data, whereas prompting offers an immediate, accessible method for LLM optimization.[15] Despite these limitations, this was, to our knowledge, the first study to directly examine how specific prompt characteristics affect LLM-based paper selection for medical MA/SRs against multiple gold-standard comparisons in a large, open-access framework.

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

5. Conclusion

Human sensitivity in abstract screening is matched by that of LLMs (86.6%) but with higher specificity in human screening.[44] In the future, LLMs could function as independent reviewers for paper screening, especially for early-stage title and abstract screening, reducing second-reviewer workload or serving as independent third reviewers. LLM-assisted screening can already reduce second-reviewer workload by 93%.[14] AI-assisted paper screening may even exceed dual-reviewer human selection in the future. Nonetheless, further optimization is required. Expectations for machine-assisted screening are higher than for human-based screening,[45] and skepticism persists about the use of LLMs in research.[48] Guideline developers have reported that alignment with established practices and transparency are the most important values in adopting automation tools in scientific research.[49] We aimed to meet part of that need by benchmarking LLM screening against posttraining MA/SRs and through the public availability of the *MA-LLM* pipeline as well as the full list of prompts used here. The *MA-LLM* pipeline may be used in future research to test any MA/SR, within and beyond medicine, and any open-source or commercial LLM. The pipeline also includes a separate option for standalone LLM-based paper screening, independent of MA/SR comparison, which was not used in the present study, but may inform future applications. Here, we used only the pipeline's option for automatically comparing LLM screening against manual MA/SR screening for the purpose of prompt engineering research. Additionally, future research may extend this pipeline beyond the NIH database and beyond title/abstract to full-text screening. We identified prompt engineering principles for LLM-based paper screening, highlighting a previously unaddressed recall/sensitivity-precision/specificity trade-off, with prompts including detailed selection criteria achieving higher F1, precision, and specificity at the expense of recall/sensitivity. Progress in developing user-friendly tools that can access and screen data from multiple literature databases, as well as implementation studies on performance are still needed until LLMs can be used reliably in paper screening for peer-reviewed journal-published MA/SRs. This work provides a step in that direction, bridging the gap between prior retrospective, dataset-specific prompt optimization and future prospective, comparison-free prompt development for new MA/SRs.

Acknowledgments. The work was conducted within the framework of the Student Network for Open Science (NOS), a volunteer-based organization that facilitated collaboration and outreach.

Author contributions. T.J.A. contributed to conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, and writing—original draft. M.D. contributed to formal analysis, software, validation, and writing—review and editing. S.A.A. and S.A.S.A. contributed to resources and writing—review and editing. E.W. contributed to data curation, formal analysis, resources, software, methodology, validation, and writing—review and editing. C.B. contributed to data curation, formal analysis, resources, and writing—review and editing. F.B., A.W., and L.K. contributed to design and writing—review and editing. C.U.C. contributed to validation and writing—review and editing. N.K. contributed to conceptualization, validation, and writing—review and editing.

Competing interest statement. C.U.C. has been a consultant and/or advisor to or has received honoraria from AbbVie, Alkermes, Allergan, Angelini, Aristo, Autobahn, Boehringer-Ingelheim, Bristol-Meyers Squibb, Cardio Diagnostics, Cerevel, CNX Therapeutics, Compass Pathways, Darnitsa, Delpor, Denovo, Draig, Eli Lilly, Eumentis Therapeutics, Gedeon Richter, GH Research, Hikma, Holmusk, IntraCellular Therapies, Jamjoom Pharma, Janssen/J&J, Karuna, LB Pharma, Lundbeck, MedInCell, MedLink Global, Merck, Mindpax, Mitsubishi Tanabe Pharmaceuticals, Maplight, Mylan, Neumora Therapeutics, Neuraxpharm, Neurocrine, Neurelis, Neurosterix, NeuShen, Neusignal Therapeutics, Newron, Noven, Novo Nordisk, Orion Pharma, Otsuka, PPD Biotech, Recognify Life Science, Recordati, Relmada, Response Pharmaeutical, Reviva, Rovi, Saladax, Sanofi, Seqirus, Servier, Sumitomo Pharma America, Sunovion, Sun Pharma, Supernus, Tabuk, Takeda, Teva, Terran, Tolmar, Vertex, Viatris, and Xenon Pharmaceuticals. He provided expert testimony for Janssen, Lundbeck, Neurocrine, and Otsuka. He served on a Data Safety Monitoring Board for Compass Pathways, IntraCellular Therapies, Relmada, Reviva, Rovi. He has received grant support from Boehringer-Ingelheim, Janssen, and Takeda. He received royalties from UpToDate and is also a stock option or stock holder of Cardio Diagnostics, Kuleon Biosciences, LB Pharmaceuticals, MedLink Global, Mindpax, Quantic, Terran.

Data Availability statement. The data and code that support the findings of this study are openly available in the *MA-LLM* repository at https://zenodo.org/records/19097181

Funding statement. The authors declare that no specific funding has been received for this article.

Supplementary material. To view supplementary material for this article, please visit http://doi.org/10.1017/rsm.2026.10093.

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

References

[1] Chen Q, Hu Y, Peng X, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nat Commun*. 2025;16(1): 3280. https://doi.org/10.1038/s41467-025-56989-2

[2] González-Márquez R, Schmidt L, Schmidt BM, Berens P, Kobak D. The landscape of biomedical research. *Patterns*. 2024;5(6): 100968. https://doi.org/10.1016/j.patter.2024.100968

[3] Sayers EW, Bolton EE, Brister JR, et al. Database resources of the National Center for biotechnology information in 2023. *Nucleic Acids Res*. 2023;51(D1): D29–D38. https://doi.org/10.1093/nar/gkac1032

[4] Wang Z, Cao L, Danek B, Jin Q, Lu Z, Sun J. Accelerating clinical evidence synthesis with large language models. *NPJ Digit Med*. 2025;8(1): 509. https://doi.org/10.1038/s41746-025-01840-7

[5] Basile C, Villaschi A, Maggioni AP. When a meta-analysis can be really useful? *Int J Cardiol*. 2025;436: 133423. https://doi.org/10.1016/j.ijcard.2025.133423

[6] 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. https://doi.org/10.1371/journal.pone.0227742

[7] 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

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

[9] Syriani E, David I, Kumar G. Screening articles for systematic reviews with ChatGPT. *J Comput Lang*. 2024;80: 101287. https://doi.org/10.1016/j.cola.2024.101287

[10] Haddaway NR, Westgate MJ. Predicting the time needed for environmental systematic reviews and systematic maps. *Conserv Biol*. 2019;33(2): 434–443. https://doi.org/10.1111/cobi.13231

[11] Cao C, Sang J, Arora R, et al. Development of prompt templates for large language model–driven screening in systematic reviews. *Ann Intern Med*. 2025;178(3): 389–401. https://doi.org/10.7326/ANNALS-24-02189

[12] Van Altena AJ, Spijker R, Olabarriaga SD. Usage of automation tools in systematic reviews. *Res Synth Methods*. 2019;10(1): 72–82. https://doi.org/10.1002/jrsm.1335

[13] Kim JK, Chua ME, Li TG, Rickard M, Lorenzo AJ. Novel AI applications in systematic review: GPT-4 assisted data extraction, analysis, review of bias. *BMJ Evid-Based Med*. https://doi.org/10.1136/bmjebm-2024-113066

[14] Delgado-Chaves FM, Jennings MJ, Atalaia A, et al. Transforming literature screening: the emerging role of large language models in systematic reviews. *Proc Natl Acad Sci*. 2025;122(2): e2411962122. https://doi.org/10.1073/pnas.2411962122

[15] Liu F, Zhou H, Gu B, et al. Application of large language models in medicine. *Nat Rev Bioeng*. 2025;3(6): 445–464. https://doi.org/10.1038/s44222-025-00279-5

[16] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8): 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

[17] 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*. Association for Computing Machinery (ACM); 2024: 262–271. https://doi.org/10.1145/3661167.3661172.

[18] Strachan JA. Designing GPT3 prompts to screen articles for systematic reviews of RCTs. *Health Policy Technol*. 2025;14(1): 100943. https://doi.org/10.1016/j.hlpt.2024.100943

[19] OpenAI. ChatGPT. 2025. https://chat.openai.com/chat

[20] Lautrup AD, Hyrup T, Schneider-Kamp A, Dahl M, Lindholt JS, Schneider-Kamp P. Heart-to-heart with ChatGPT: the impact of patients consulting AI for cardiovascular health advice. *Open Heart*. 2023;10(2): e002455. https://doi.org/10.1136/openhrt-2023-002455

[21] Nguyen D, MacKenzie A, Kim YH. Encouragement vs. liability: how prompt engineering influences ChatGPT-4's radiology exam performance. *Clin Imaging*. 2024;115: 110276. https://doi.org/10.1016/j.clinimag.2024.110276

[22] Web of Science Journal Info: Impact factor, Quartile, Ranking. https://wos-journal.info/

[23] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021; n71. https://doi.org/10.1136/bmj.n71

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

[25] Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *Npj Digit Med*. 2024;7(1): 41. https://doi.org/10.1038/s41746-024-01029-4

[26] Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. *Patterns*. 2025;6(6): 101260. https://doi.org/10.1016/j.patter.2025.101260

[27] Millard J. markovify: A simple, extensible Markov chain generator. 2015. https://github.com/jsvine/markovify

[28] Van Rossum G, Drake FL . Python. 2009. https://docs.python.org/3/reference/

[29] Meta AI. Meta Llama 3-8B. 2024. https://huggingface.co/meta-llama/Meta-Llama-3-8B

[30] Meta AI. Meta Llama 3-70B. 2024. https://huggingface.co/meta-llama/Meta-Llama-3-70B

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press

[31] Mistral AI. Mistral Large 2407. 2024. https://mistral.ai/news/mistral-large-2407
[32] Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PMM. Systematic reviews of diagnostic test accuracy. *Ann Intern Med*. 2008;149(12): 889–897. https://doi.org/10.7326/0003-4819-149-12-200812160-00008
[33] Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016: i6. https://doi.org/10.1136/bmj.i6
[34] Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007;115(5): 654–657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929
[35] Powers D. *Evaluation: From Precision, Recall and F-Factor to ROC*. Informedness, Markedness & Correlation.
[36] Brauer M, Curtin JJ. Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychol Methods*. 2018;23(3): 389–411. https://doi.org/10.1037/met0000159
[37] R Core Team. R: A Language and Environment for Statistical Computing. 2024. https://www.R-project.org/
[38] Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using **lme4**. *J Stat Softw*. 2015;67(1). https://doi.org/10.18637/jss.v067.i01
[39] Kuznetsova A, Brockhoff PB, Christensen RHB. **lmerTest** package: tests in linear mixed effects models. *J Stat Softw*. 2017;82(13). https://doi.org/10.18637/jss.v082.i13
[40] Geta ET, Terefa DR, Hailu WB, et al. Effectiveness of shared decision-making for glycaemic control among type 2 diabetes mellitus adult patients: a systematic review and meta-analysis. *PLoS One*. 2024;19(7): e0306296. https://doi.org/10.1371/journal.pone.0306296
[41] Santmartí-Oliver M, Jorba-García A, Moya-Martínez T, de-la-Rosa-Gay C, Camps-Font O. Safety and accuracy of guided Interradicular Miniscrew insertion: a systematic review and meta-analysis. *J Clin Med*. 2024;13(24): 7697. https://doi.org/10.3390/jcm13247697
[42] Yu X, Tan Q, Li J, Shi Y, Chen L. Elective high frequency oscillatory ventilation versus conventional mechanical ventilation on the chronic lung disease or death in preterm infants administered surfactant: a systematic review and meta-analysis. *J Perinatol*. 2025;45(1): 77–84. https://doi.org/10.1038/s41372-024-02185-x
[43] Yang X, Fang S, Lyu W, et al. Vortioxetine for depression in adults: a systematic review and dose–response meta-analysis of randomized controlled trials. *Psychiatry Clin Neurosci*. 2024;78(9): 536–545. https://doi.org/10.1111/pcn.13709
[44] Gartlehner G, Affengruber L, Titscher V, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *J Clin Epidemiol*. 2020;121: 20–28. https://doi.org/10.1016/j.jclinepi.2020.01.005
[45] Hanegraaf P, Wondimu A, Mosselman JJ, et al. Inter-reviewer reliability of human literature reviewing and implications for the introduction of machine-assisted systematic reviews: a mixed-methods review. *BMJ Open*. 2024;14(3): e076912. https://doi.org/10.1136/bmjopen-2023-076912
[46] Stoll CRT, Izadi S, Fowler S, Green P, Suls J, Colditz GA. The value of a second reviewer for study selection in systematic reviews. *Res Synth Methods*. 2019;10(4): 539–545. https://doi.org/10.1002/jrsm.1369
[47] Zhang G, Xu Z, Jin Q, et al. Leveraging long context in retrieval augmented language models for medical question answering. *Npj Digit Med*. 2025;8(1): 239. https://doi.org/10.1038/s41746-025-01651-w
[48] Mishra T, Sutanto E, Rossanti R, et al. Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. *Sci Rep*. 2024;14(1): 31672. https://doi.org/10.1038/s41598-024-81370-6
[49] Arno A, Elliott J, Wallace B, Turner T, Thomas J. The views of health guideline developers on the use of automation in health evidence synthesis. *Syst Rev*. 2021;10(1): 16. https://doi.org/10.1186/s13643-020-01569-2

Cite this article: Adam TJ, Abosabie SAS, Dittmer M, Wolf E, Abosabie SA, Behnke C, Baier F, Weickmann A, Köser L, Correll CU, Rutsch N. Prompt engineering of large language models for paper screening in medical meta-analyses and systematic reviews: A prospective comparative study. *Research Synthesis Methods*. 2026;17: 939–956. https://doi.org/10.1017/rsm.2026.10093

https://doi.org/10.1017/rsm.2026.10093 Published online by Cambridge University Press