International Journal of Technology Assessment in Health Care

www.cambridge.org/thc

Method

Cite this article: Qiu T, Jiang H, Chen MS, Lin T (2026). A prompt-based tutorial for large language model–assisted screening in systematic reviews and meta-analyses. *International Journal of Technology Assessment in Health Care*, **42**(1), e83, 1–9
https://doi.org/10.1017/S0266462326104073

Received: 07 November 2025 UTC
Revised: 13 April 2026 UTC
Accepted: 19 July 2026 UTC

Keywords:
large language models; ChatGPT; systematic review; meta-analysis; prompt engineering

Corresponding author:
Tao Lin;
Email: tao.lin@pennmedicine.upenn.edu

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

A prompt-based tutorial for large language model–assisted screening in systematic reviews and meta-analyses

Tongyu Qiu[1], Huangqi Jiang[2], Mark S. Chen[3] and Tao Lin[4]

[1]Department of Social and Behavioral Sciences, School of Public Health, Yale University, USA; [2]Department of Psychological Sciences, Case Western Reserve University Hospital, USA; [3]Department of Psychology, Harvard University, USA and [4]Department of Psychiatry, Center for Treatment and Study of Anxiety, University of Pennsylvania, USA

Abstract

Objectives: Title/abstract and full-text screening are among the most time-consuming stages of systematic reviews. Large language models (LLMs) such as ChatGPT can assist in screening, but prior evaluations have largely focused on abstract-level decisions and require coding or application programming interfaces (API) access, limiting practical application. This study developed and evaluated a structured, prompt-based approach that enables LLMs to perform both abstract and full-text screening without programming or API integration.

Methods: Using datasets from three completed meta-analyses, we implemented a stepwise training framework involving comprehension checks, criterion-specific feedback, and iterative prompt refinement. Study 1 optimized file format and batch size using 1,000 abstracts and 50 full texts. Studies 2 and 3 validated the approach using independent datasets on teletherapy versus in-person therapy (1,321 abstracts, 82 full texts) and emotion regulation and mental health (1,500 abstracts, 129 full texts), respectively. Human reviewers' decisions served as the reference standard.

Results: In Study 1, the LLM achieved 98 percent sensitivity and 80.6 percent specificity at abstract screening, with batches of 50 abstracts in plain-text format performing optimally. In Study 2, sensitivity and specificity were 100 percent and 85.6 percent for abstracts, with 82.1 percent full-text accuracy and all eligible studies retained. In Study 3, abstract screening achieved 100 percent sensitivity and 93 percent specificity, while full-text screening achieved 96.2 percent sensitivity and 90.3 percent specificity.

Conclusions: A structured, prompt-based approach enabled LLMs to achieve high sensitivity and specificity across both screening stages without coding or APIs. With human oversight, this accessible workflow can reduce reviewer burden and accelerate evidence synthesis.

Introduction

Systematic reviews and meta-analyses underpin evidence-based practice, synthesizing primary research to inform clinical guidelines, health policy, and future research priorities (1). However, these studies require intensive labor, particularly during the title, abstract, and full-text screening phases (2;3). Reviewers need to process thousands of records to identify the handful of studies that meet eligibility criteria (4), and screening alone can consume up to 20 percent of the total time invested in a review (5). This manual process is not only time-consuming but also resource-intensive (6), limiting the efficiency and scalability of evidence synthesis.

Recent advances in artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT, present new opportunities to enhance the efficiency of systematic review workflows. Conventional machine learning-based screening tools approach this task by training supervised classifiers to distinguish between eligible and ineligible studies. This process generally requires researchers to create a labeled dataset of inclusion and exclusion decisions, design features from the text (for example, word frequencies, key phrases, or topic models), and program and maintain algorithms such as logistic regression or support vector machines to make predictions (7;8). While these systems can reduce workload once optimized, they require substantial upfront investment and technical expertise. In practice, researchers must devote considerable time to manually creating large, labeled datasets and applying complex feature-engineering techniques, both of which demand specialized skills and limit widespread adoption (9;10).

In contrast, LLMs can interpret detailed natural language instructions ("prompts") and apply complex inclusion and exclusion criteria without any formal retraining (11). This prompt-driven approach confers several advantages over conventional automated screening pipelines. First, researchers without computational training can access these capabilities through user-friendly interfaces, lowering the barrier to adoption (12). Second, the approach is readily adaptable. By

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 09:23:08, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0266462326104073

simply modifying or refining the prompt, reviewers can adjust the agent's decision-making criteria for different research questions, study designs, or populations, without rebuilding or retraining the agent (13). Third, LLMs can apply nuanced eligibility criteria that are difficult to encode in rule-based systems (14), including interpreting contextual cues, managing exceptions, and integrating multiple conditions into a single decision. In a review restricted to adults, for instance, a model can recognize that a study of both children and adults may still be eligible when outcomes are reported separately by age group. Fourth, LLMs accept diverse inputs such as abstracts, titles, full-text PDFs, and plain-text files (.txt), while maintaining consistent reasoning across documents. Given these advantages, LLMs are promising tools for title/abstract screening and full-text review in systematic reviews and meta-analyses, especially for research teams having limited access to specialized programming infrastructure.

Research evaluating agent-assisted screening has shown promise in reducing reviewer workload while maintaining moderate-to-high sensitivity (15;16). For example, research demonstrated that LLMs could achieve high accuracy and sensitivity in title/abstract screening across multiple domains (17). Yet their use in systematic reviews and meta-analyses remains at an early stage, and several gaps limit practical uptake. First, questions persist about the reproducibility of decisions across sessions and models, and about best practices for embedding LLMs in established systematic review workflows in a way that preserves transparency and methodological rigor (9;10). The reviews that have incorporated LLMs seldom report how the model was trained or guided, and no standardized procedure yet exists, raising concerns about screening accuracy. Second, these studies have primarily focused on abstract screening, leaving the more complex full-text stage underexplored (11;18). Third, many workflows relied heavily on application programming interfaces (APIs) and coding to automate the retrieval and processing of full-text articles. APIs are often necessary because LLMs cannot efficiently handle very large batches of documents and have limited memory for retaining earlier inputs, making automated pipelines essential for scaling systematic reviews. However, the use of APIs introduces additional barriers, as it requires programming expertise, technical infrastructure, and financial investment, which limits accessibility for many research teams.

To address these gaps, the present study aimed to develop and validate a prompt-based tutorial for using ChatGPT, one of the most widely used LLMs, for abstract and full-text screening in systematic review and meta-analysis. A prompt-based approach was selected because it eliminates the need for programming, custom APIs, or fine-tuning, lowering the technical and financial barriers that often prevent research teams from adopting AI tools. Unlike API-driven or code-dependent pipelines, prompt-based workflows can be implemented entirely within user-friendly interfaces, keeping them accessible to non-technical reviewers while preserving transparency, reproducibility, and adaptability across different review contexts. Using a completed meta-analysis that examines psychotherapy formats for anxiety disorders as a test dataset, we first developed an iterative, feedback-informed prompting approach to improve the agent's comprehension and application of complex eligibility criteria and to identify practical parameters such as file formats and batch sizes. We then examined the accuracy of this approach in article screening by comparing its screening decisions with those of human reviewers at both the abstract and full-text stages in another meta-analysis project.

Study 1: development of a prompt-based tutorial

Methods

Overview of original meta-analysis and setup

To develop a prompt-based screening tutorial for abstract screening and full-text screening, we used a dataset from a recently completed meta-analysis that compared the efficacy of various psychotherapy formats in adults with anxiety disorders. This dataset provided a testbed for iteratively refining and evaluating the screening prompts. In the original review, two independent reviewers screened the titles and abstracts of all database-identified records against pre-specified eligibility criteria. Articles judged potentially eligible advanced to full-text review, again conducted independently by two reviewers. Discrepancies between coders were resolved through discussion, and a third reviewer was consulted in cases where consensus could not be reached. This approach followed best practices for systematic review procedures as outlined in prior methodological work (9;10). For the present study, we selected a subset of 1,000 articles from this screened pool, comprising 900 irrelevant articles (excluded at the abstract screening stage), 50 articles excluded at the full-text screening stage, and 50 studies that met all eligibility criteria and were included in the meta-analysis.

Human screening decisions were compiled in .csv format to serve as the reference standard for evaluating agent performance. The inclusion and exclusion criteria were documented in a PDF file and uploaded into a new ChatGPT project folder (see Supplementary Table S1 for the full list of inclusion and exclusion criteria). We used PDF format for reliable ingestion, as prior attempts to upload Excel spreadsheets frequently resulted in recognition errors or misaligned formatting within the agent interface.

Preparing the agent for screening

Figure 1 presents the full training procedure. To ensure the agent could both process documents completely and accurately apply the screening criteria, we implemented two preparatory safeguards before initiating the workflow. First, to manage token limitations, we restricted each batch to 5–6 uploaded files per session. Token limitations refer to the maximum amount of text, measured in tokens (units of words or word parts), that the agent can process in a single session. When this limit is exceeded, earlier input may be truncated or disregarded. After each batch, we refreshed the session in a new ChatGPT window, resetting the token count so that subsequent files could be processed without loss of information. Second, to confirm that the agent had correctly interpreted the eligibility criteria, we initiated the training process with a comprehension check. This step served as a quality control measure to assess the agent's initial understanding of the uploaded criteria.

After uploading the criteria document (in PDF format), we prompted the agent to read and interpret the content by asking it to summarize the criteria in its own words. An example of the prompt used is:

> *You will be asked to screen abstracts based on this uploaded PDF file outlining inclusion and exclusion criteria. Please summarize the criteria in your own words.*

This step confirmed whether the agent could accurately extract and paraphrase key rules, such as eligibility based on population, intervention type, study design, and outcome reporting. We review each summary completeness and accuracy and provide direct feedback whenever the agent omitted a detail, misinterpreted a criterion, or introduced an inconsistency. When needed, we revised

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 09:23:08, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0266462326104073

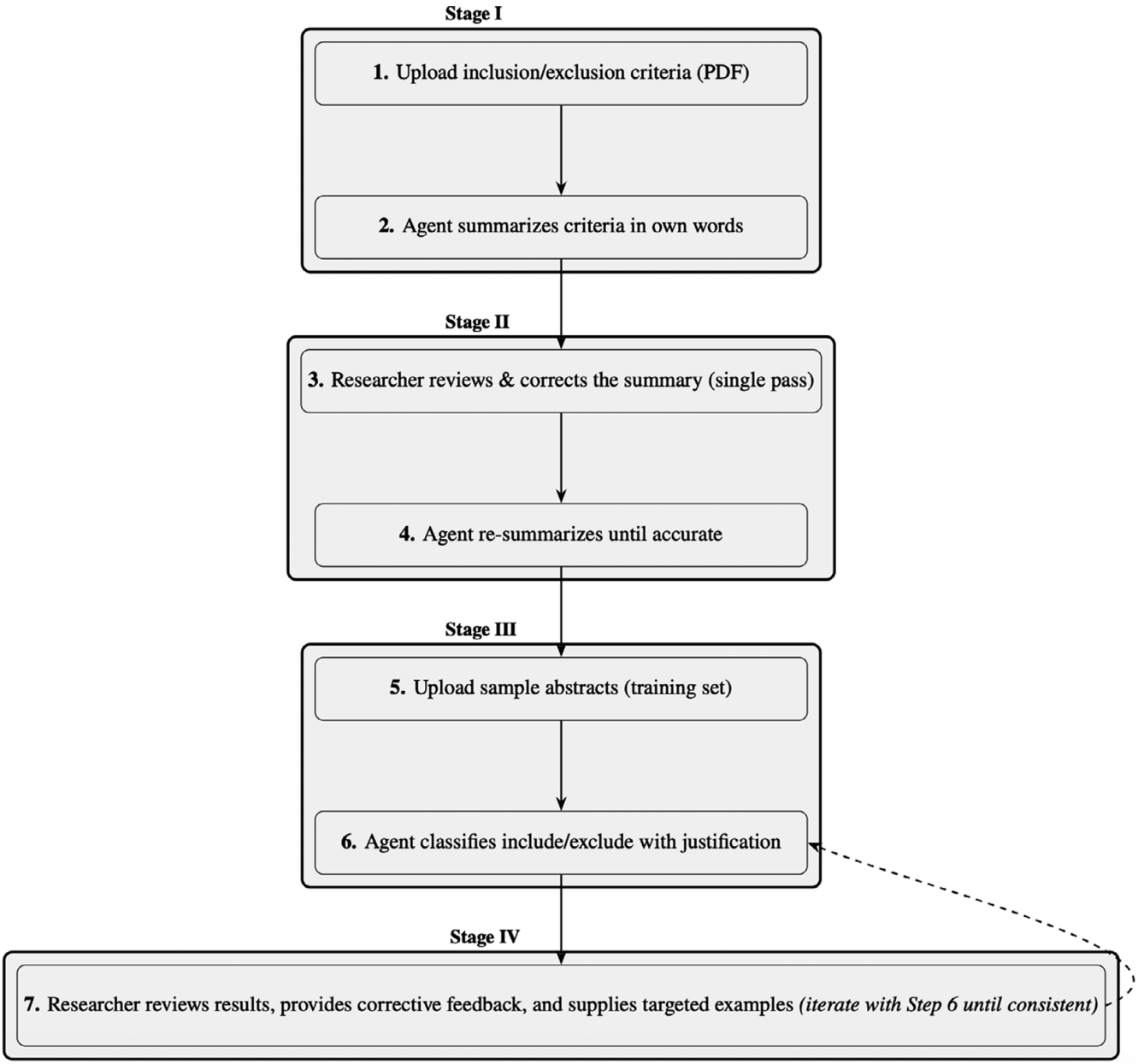


Figure 1. Training process in Studies 1 and 2.

the prompts by (a) explicitly restating overlooked rules, (b) adding clarifying examples of how a criterion should be applied, or (c) restructuring the instructions into shorter, stepwise questions to reduce ambiguity. For instance, if the agent failed to distinguish adults from adolescents, we modified the prompt to say:

> *Summarize the inclusion and exclusion criteria. Be sure to specify the required age group, and note that studies involving adolescents should be excluded unless adult data are reported separately.*

This iterative feedback process ensured that the agent developed a sound foundational understanding before proceeding to screening tasks.

Next, to assess whether the agent could apply the criteria, we provided a series of training examples that illustrated exclusion scenarios. For each major criterion (for example, population age, intervention type, study design), we selected abstracts that clearly violated that rule. The agent was asked whether the study should be excluded and to justify its decision. Because a single study could fail multiple criteria simultaneously, we checked whether the agent cited at least one correct reason for exclusion, even if additional criteria also applied. Our training was therefore focused on teaching the agent to recognize valid exclusion grounds rather than requiring it to enumerate every possible reason. A typical prompt was:

> *Based on the uploaded criteria, should the following study be included or excluded? Please explain your reasoning.*

We reviewed each response for correctness and corrected errors directly. If the agent incorrectly included or excluded a study, direct feedback was provided. For example, if the agent recommended inclusion for a study with nonadult participants when the criteria

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 09:23:08, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0266462326104073

required adult samples, the user would reply with a correction such as:

> *This study should be excluded because the participants are teenagers. The inclusion criteria require adult participants. See the sentence in the abstract that says: 'Participants were adolescents aged 13–17.' This criterion was not met.*

We then supplied three to five further exemplars targeting the same criterion, allowing the agent to reinforce and consolidate the rule. This iterative process was repeated for each major criterion until the agent demonstrated consistent accuracy. This iterative refinement process is not merely about correcting errors but about aligning the model's internal decision-making threshold with the specific nuance of the inclusion criteria. By focusing on criterion-specific corrections and practice, we improved the reliability of the agent's screening decisions across varied scenarios.

Agent test

After the initial training and comprehension check, we evaluated the agent's performance through a series of controlled test batches. Each test consisted of a randomized set of study articles presented in a *single plain-text file* (.txt). To identify the most reliable format, we evaluated three options: plain-text (.txt), PDF, and Word (.docx), using the same sample of ten articles uploaded in each format prior to screening. Classification accuracy was the primary outcome, assessed by comparing the agent's inclusion and exclusion decisions against human reviewer judgments. As a secondary check, we also verified whether the agent's stated reasoning matched its decision and accurately reflected the content of the article, rather than reflecting hallucinated or unsupported justifications. Plain-text format demonstrated consistent and reliable parsing without layout distortion, ensuring that the full content of each article was accessible to the agent. PDF files occasionally produced layout distortions and formatting inconsistencies, particularly when articles contained multi-column layouts or embedded images, which sometimes caused misaligned text extraction (19). Word (.docx) files were associated with formatting recognition errors and inconsistent rendering of tables and special characters within the agent interface (20). On the basis of these observations, plain-text format was adopted as the standard input format for all screening tasks.

Each file included a mixture of abstracts previously excluded at the abstract or full-text stage, as well as studies that were ultimately included. A standard evaluation prompt was:

> *Please review the titles and abstracts of the following [N] articles based on the criteria PDF in the project folder. For each article, indicate 'Include' or 'Exclude' and provide a brief reason based on the criteria.*

To determine the optimal batch size, we ran internal tests across varying input lengths. The same pool of articles, containing a mixture of excluded and included studies, was used across all conditions; articles were simply reorganized into batches of different sizes (20, 50, 100, and 200 articles per batch) for each test. Sensitivity and specificity were calculated for each batch size configuration to identify the optimal balance between input length and screening accuracy.

Technical implementation

The screening process was conducted using the GPT-4o model. No API, fine-tuning, or external plugins were used. Prompt inputs and outputs were managed manually through the web-based interface. This manual setup ensured that the process was transparent and reproducible without relying on backend automation or custom coding.

Results

Internal testing of batch size configurations showed that performance declined when more than 50 articles were screened at once. Batches of 20 articles produced consistently high sensitivity and specificity but were impractical, given the large number of sessions required to screen thousands of records. Batches of 50 articles maintained high and stable performance across multiple test runs and were identified as the most efficient balance between workload and accuracy. Performance declined noticeably at 100 articles, with greater variability in both sensitivity and specificity across batches, likely reflecting the agent's reduced attention over longer inputs due to the lost-in-the-middle phenomenon and token limitations. At 200 articles, specificity collapsed substantially, indicating that the agent began defaulting to broad inclusion rather than applying the criteria discriminately. On this basis, a batch size of 50 articles was adopted as the operational standard for abstract screening (see Supplementary Table S2). The agent recommended inclusion for 141 of the 900 irrelevant studies, 43 of the 50 studies later excluded at full-text, and 49 of the 50 studies that met final eligibility criteria. These inclusion patterns yielded a sensitivity of 98. percent and a specificity of 80.6 percent (See Table 1). These results guided the final refinement of the prompts prior to external validation.

Study 2: validation of the prompt-based screening tutorial

To evaluate the performance of the screening approach, we applied it to a published meta-analysis that examined whether synchronous teletherapy achieves similar clinical outcomes and client retention as in-person therapy (21). The evaluation consisted of two phases: (1) title and abstract screening and (2) full-text screening. Study 1 focused on abstract screening to establish and refine the prompting strategy, while Study 2 extended the evaluation to full-text screening to determine whether the approach could be applied effectively to this more complex stage.

Methods

Dataset preparation for agent evaluation

To minimize order effects and reduce potential bias, all screened articles were pooled and redistributed into mixed batches before agent evaluation. A total of 1,219 articles were drawn from three outcome categories based on human reviewer decisions: studies excluded at the title and abstract screening stage (n = 1,137), studies

Table 1. Agent inclusion and exclusion decisions compared to the human reviewer ground truth during the developmental stage

	Agent: include	Agent: exclude	Total
Human: eligible (n = 50)	49	1	**50**
Human: ineligible (n = 950)	184	766	**950**
Total	233	767	**1,000**
Sensitivity	98.0%		
Specificity		80.6%	
Overall accuracy			81.5%

Note: Rows reflect human reviewer ground truth labels; columns reflect agent decisions.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 09:23:08, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0266462326104073

that advanced to but were excluded at full-text screening ($n = 62$), and studies that were ultimately included in the final review ($n = 20$).

Agent training at abstract screening phase

The training procedure closely followed the steps established in Study 1 (see Figure 1). First, the inclusion and exclusion criteria were uploaded in PDF format. The agent was instructed to read the document and summarize the criteria in its own words. Iterative feedback was provided until the summary was complete and accurate, thereby ensuring that the agent had an adequate foundation before applying the rules to specific studies.

Following this comprehension check, we presented the agent with ten randomly selected study titles and abstracts. For each article, the agent was prompted to decide whether the study should be included or excluded and to provide a justification based on the uploaded criteria. Misclassifications were corrected through explicit feedback that referenced the relevant criterion. When necessary, additional targeted examples were supplied to reinforce specific rules that the agent had misapplied. In total, 25 abstracts were used across training, with at least two to three examples per major criterion. Training was considered complete once the agent demonstrated consistent accuracy across all presented examples and no further misclassifications occurred for two consecutive rounds of testing.

Agent evaluation at abstract screening phases

At this phase, we evaluated the agent's performance on screening titles and abstracts against the pre-specified eligibility criteria. To standardize inputs and minimize formatting errors, all articles were exported as plain-text (.txt) files in batches of 50, a limit informed by preliminary testing in Study 1. For each batch, the agent received a structured prompt instructing it to: (a) apply the uploaded inclusion and exclusion criteria, (b) return a binary decision ("Include" or "Exclude") for each article, and (c) provide a brief, criteria-based justification for its decision. The agent's output was then exported and archived for comparison with human reviewer judgments.

To quantify performance, the agent's decisions were mapped onto the final human screening results. A true positive was defined as a study included by both the agent and human reviewers, while a true negative referred to a study excluded by both. False positives were studies flagged for inclusion by the agent but excluded by human reviewers, and false negatives were studies excluded by the agent but included by humans. Sensitivity and specificity were calculated from these classifications to evaluate the accuracy of the agent's decisions at the abstract stage. All studies flagged as "Include" by the agent were advanced to the subsequent full-text screening phase.

Agent evaluation at full-text screening phases

At the full-text stage, we assessed the agent's ability to apply the eligibility criteria to full-text. All studies retained at the abstract stage were retrieved in PDF format and uploaded for review. As with abstract screening, we conducted internal tests to determine the optimal batch size for full-text screening. These tests showed that when more than three articles were uploaded at once, the agent frequently produced incomplete or partial decisions. In contrast, batches of three articles were processed reliably and with consistent accuracy. On this basis, we set three full-text articles per batch as the operational default.

For each batch, the agent was provided with a structured prompt instructing it to: (a) apply the same inclusion and exclusion criteria used during abstract screening, (b) return a binary decision ("Include" or "Exclude") for each article, and (c) provide verbatim quotations from the text to justify its decision. Requiring direct evidence served to increase transparency and allowed human reviewers to verify the basis of each classification. For example, the prompt instructed the agent:

> *"Assess whether this full-text article meets the inclusion criteria for this systematic review. If you decide to include or exclude, provide direct quotations from the article to justify your decision.*

The agent's outputs were then exported and archived for comparison with human reviewer judgments. To quantify performance, the agent's inclusion decisions were mapped onto the final human-coded labels, and overall accuracy was calculated. Accuracy reflected the proportion of studies for which the agent's decision matched the final human judgment, providing a single metric of agreement at the full-text stage.

Results

At the abstract and title screening stage, the agent selected a total of 193 articles for full-text review. Of these, 173 were ineligible and 20 were correctly identified as eligible for inclusion, yielding a sensitivity of 100 percent and a specificity of 85.6 percent (See Table 2 and Figure 2).

The 193 retained studies then proceeded to full-text screening, where the agent recommended inclusion for 25 studies. Critically, all 20 studies that met the final eligibility criteria according to human reviewers were correctly retained. This yielded a sensitivity of 100 percent and a specificity of 97.1 percent, demonstrating that the agent successfully identified all eligible studies while accurately excluding the majority of ineligible ones. When compared against the final human inclusion list, this corresponded to an overall accuracy of 97.4 percent, indicating that the agent was able to approximate human screening decisions at the full-text level while maintaining high sensitivity.

Study 3: cross-domain validation in a nonintervention observational review

Methods

Overview of original meta-analysis and dataset

To evaluate whether the agent-assisted screening workflow generalizes beyond intervention- and RCT-focused reviews and psychotherapy-specific content, we conducted a third validation using a non-RCT design systematic review and meta-analysis on emotion regulation and mental health across cultures (22). This meta-analysis differs from those used in Studies 1 and 2 in two important aspects. First, instead of examining intervention efficacy, this review focused on the associations between emotion regulation strategies and a range of mental health outcomes across cultures, a distinct topic. Second, and more critically, the study design type is fundamentally different: while the two meta-analyses used in Studies 1 and 2 focused on RCTs, this meta-analysis targeted observational, correlational studies that measure naturally occurring relationships between emotion regulation and mental health and excluded studies involving experimental manipulation of emotion regulation, purely intervention designs, and intensively longitudinal methods. This represents a methodologically distinct evidence base with considerably greater heterogeneity in study design, population characteristics, cultural context, and outcome operationalization than the RCT-based intervention reviews used in Studies 1 and 2.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 09:23:08, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0266462326104073

Table 2. Agent inclusion and exclusion decisions compared with human reviewer ground truth during the validation stage (Study 2: RCT- and Intervention-Based Meta-Analysis)

	Agent: include	Agent: exclude	Total
Title and abstract screening (*n* = 1,219)			
Human: Eligible (*n* = 20)	20	0	**20**
Human: Ineligible (*n* = 1,199)	173	1,026	**1,199**
Total	193	1,026	**1,219**
Sensitivity	100.0%		
Specificity		85.6%	
Overall accuracy			85.8%
Full-text review (*n* = 193)			
Human: eligible (*n* = 20)	20	0	**20**
Human: ineligible (*n* = 173)	5	168	**173**
Total	25	168	**193**
Sensitivity	100.0%		
Specificity		97.1%	
Overall accuracy			97.4%

Note: Rows reflect human reviewer ground truth labels; columns reflect agent decisions.

To test the generalizability of the workflow, we drew a random sample of 1,500 articles from the pool identified through the original database search. This sample was stratified to include both eligible and ineligible studies consistent with the overall composition of the pool, preserving the realistic class imbalance encountered in abstract screening and full-text review. Human reviewer decisions on these records served as the reference standard.

Agent training and evaluation

The same prompt-based training procedure established in Studies 1 and 2 was applied without modification (Figure 1). The inclusion and exclusion criteria were uploaded in PDF format, and the agent completed a comprehension check before screening. Iterative corrective feedback was provided until the agent demonstrated accurate and consistent interpretation of all major criteria, including nuanced eligibility rules related to cultural subgroup reporting requirements, instrument validation standards, and sample exclusion logic. Abstract screening was conducted using batches of 50 plain-text articles, consistent with Studies 1 and 2. Full-text screening was conducted using batches of three articles in PDF format, as established in Study 2. Sensitivity, specificity, and overall accuracy were calculated at both stages using human reviewer decisions as the reference standard.

Results

Title and abstract screening

Of the 1,500 randomly sampled records (eligible = 26; Ineligible = 1,474), the agent correctly identified all 26 eligible studies, achieving 100 percent sensitivity, indicating that no eligible studies were missed. Of the ineligible records, 1,371 were correctly excluded and 103 were advanced to full-text review, yielding 93 percent specificity and 93.1 percent overall accuracy (see Table 3).

Full-text review

Of the 129 records advanced to full-text review (eligible = 26; Ineligible = 103), the agent correctly identified 25 eligible studies, achieving 96.2 percent sensitivity. Of the ineligible records, 93 were

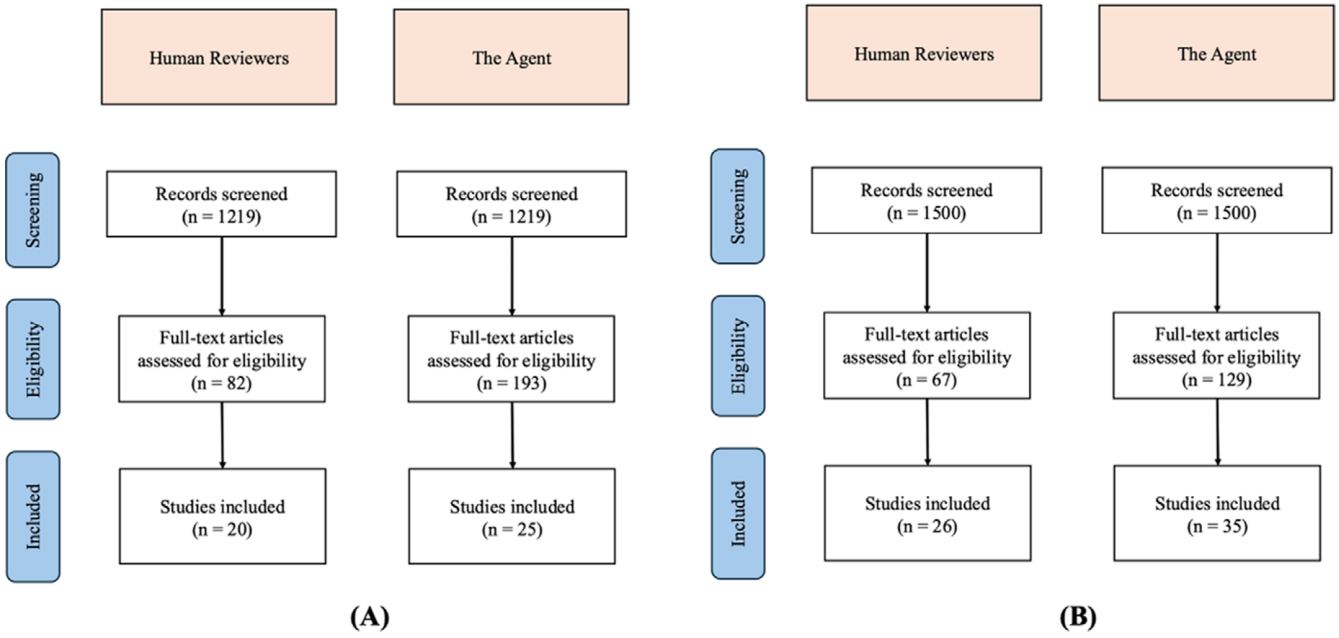


Figure 2. Comparison of the study selection process between human reviewers and the agent across the stages of screening, eligibility, and inclusion. (A) Study 2: RCT- and intervention-based meta-analysis. (B) Study 3: Non-RCT and nonintervention-based meta-analysis.
Note: In Study 2, of the twenty-five studies the agent recommended for inclusion, twenty matched the human reviewers' final inclusion list (true positives), and five did not (false positives). All twenty studies ultimately included by human reviewers were correctly identified by the agent (sensitivity = 100%). In Study 3, of the thirty-five studies the agent recommended for inclusion, twenty-five matched the human reviewers' final inclusion list (true positives), and ten did not (false positives). One eligible study was not identified by the agent at the full-text stage (false negative), resulting in a sensitivity of 96.2%.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 09:23:08, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0266462326104073

Table 3. Agent inclusion and exclusion decisions compared with human reviewer ground truth during the validation stage (Study 3: Non-RCT and Nonintervention-Based Meta-Analysis)

	Agent: include	Agent: exclude	Total
Title and abstract screening (n = 1,500)			
Human: eligible (*n* = 26)	26	0	**26**
Human: ineligible (*n* = 1,474)	103	1,371	**1,474**
Total	129	1,371	**1,500**
Sensitivity	100.0%		
Specificity		93.0%	
Overall accuracy			93.1%
Full-text review (*n* = 129)			
Human: eligible (*n* = 26)	25	1	**26**
Human: ineligible (*n* = 103)	10	93	**103**
Total	35	94	**129**
Sensitivity	96.2%		
Specificity		90.3%	
Overall accuracy			91.5%

Note: Rows reflect human reviewer ground truth labels; columns reflect agent decisions.

correctly excluded, yielding 90.3 percent specificity and 91.5 percent overall accuracy (see Table 3).

Discussion

These results closely replicate the performance patterns observed in Studies 1 and 2. The replication of near-perfect sensitivity at the abstract stage, with zero eligible studies missed, across a nonintervention, observational evidence base with more complex and heterogeneous eligibility criteria provides evidence that the prompt-based workflow generalizes across both topic domains and study design types.

Data handling and ethical considerations

The data used in this study were derived from manuscripts under review and published meta-analyses conducted by the research team. No personally identifiable information or unpublished primary data were used. The project did not involve human participants or clinical interventions and was therefore exempt from institutional ethics review. The use of AI for screening was limited to title, abstract, and full-text classification tasks and did not involve any form of automated decision-making with clinical or regulatory implications. Prompts, training examples, and evaluation tutorials have been documented and are available in Supplementary Material S2 to support transparency and reproducibility. Researchers who wish to adopt this workflow for full-text screening should be aware that uploading paywalled or proprietary full-text articles into AI platforms may raise copyright concerns depending on their jurisdiction and institutional policies. We strongly encourage researchers to consult their institutional copyright guidelines and seek appropriate legal or library advice before uploading any copyrighted materials to AI platforms to ensure compliance with applicable laws and publisher agreements. A public version of the developed GPT (*LLM Screener for Systematic Reviews*) is available for replication (see Supplementary Material S3).

General discussion

This study developed and validated a prompt-based tutorial for using an LLM, specifically ChatGPT, to conduct both abstract and full-text screening for systematic reviews and meta-analyses. Across three independent datasets, we showed that an iterative, feedback-driven prompting approach can deliver high sensitivity and satisfactory specificity across both screening stages, without the need for API access, model fine-tuning, or specialized programming skills. In Study 1, the developmental dataset, the agent achieved 98.0 percent sensitivity and 80.6 percent specificity at abstract screening; in Study 2, an RCT-based teletherapy review, performance rose to 100 percent sensitivity and 85.6 percent specificity, retaining all eligible studies while limiting false positives. At the more nuanced full-text stage in Study 2, the agent reached 97.4 percent overall accuracy relative to human reviewers, correctly identifying every eligible study. Study 3 extended this validation to a methodologically distinct context, an observational, correlational meta-analysis on emotion regulation and mental health across cultures, where the agent again achieved 100 percent sensitivity at abstract screening and 96.2 percent sensitivity at full-text review, with no eligible studies missed at the abstract stage. The consistency of these results across three independent datasets, two clinical domains, and two fundamentally different evidence types strengthens confidence in the generalizability and robustness of the proposed workflow.

Our findings align with and build upon prior research demonstrating that LLMs can function as effective assistants in the screening stages of evidence synthesis (23;24). Earlier studies have consistently shown that LLMs achieve high sensitivity during title and abstract screening (23;24), often retaining nearly all eligible studies while substantially reducing the number of records requiring manual review. Our results replicated this pattern, with the approach consistently identifying all eligible studies at the abstract stage, ensuring that no eligible studies were prematurely excluded. However, this came at the cost of lower specificity: the agent advanced several ineligible studies to full-text review in all three datasets, reflecting a deliberate bias toward inclusion when abstract information is insufficient. For example, five studies incorrectly flagged as eligible required additional review in Study 2, reflecting a trade-off between inclusivity and efficiency. Interestingly, despite the agent's tendency to retain more studies at the title/abstract phase than human reviewers, the final number of eligible studies after full-text screening was nearly identical. This suggests that LLMs may have difficulty making accurate inclusion decisions with limited information provided in the title and abstract.

Importantly, in both Study 2 and Study 3, the agent demonstrated substantially improved specificity at the full-text stage compared to abstract screening, converging toward human-level discriminative accuracy when provided with the complete evidence base. This pattern suggests that LLMs may compensate for information scarcity in abstracts by defaulting toward inclusion, while demonstrating considerably greater discriminative capacity when full-text content is available. This convergence has practical implications for the design of hybrid human–AI screening workflows: a high-sensitivity LLM pass at the abstract stage can safely reduce the volume of records requiring human attention, while a modest human check at the full-text stage is sufficient to correct residual false positives without sacrificing recall. Taken together, these

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 09:23:08, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0266462326104073

findings underscore both the promise and the limitations of LLM-assisted screening: future researchers can rely on such tools to safeguard sensitivity but must plan for continued human oversight to correct misclassifications and ensure methodological rigor.

Our study advanced this literature in several important ways. First, we systematically evaluated agent performance at the full-text stage, which has received relatively little attention in prior work. By demonstrating that an agent can approximate human-level accuracy in identifying detailed participant characteristics, intervention tutorials, and outcome measures during full-text screening, we provides novel evidence that these tools can support the most resource-intensive phase of systematic reviews.

Second, a central factor in this performance was a deliberate prompt design that required the agent not only to render an inclusion or exclusion decision but also to supply verbatim quotations from the source text in support. This requirement served two functions: it increased transparency by allowing human reviewers to verify the basis for each decision, and it encouraged the agent to engage in targeted retrieval of relevant information rather than relying on general impressions of the text. By prompting the agent to cite specific passages, we also mitigated one of the known risks of automated screening, overreliance on implicit reasoning without traceable evidence (25), which can undermine trust in the agent's outputs. This design choice is consistent with recommendations from previous AI-in-evidence-synthesis research (11), which emphasizes that explainability and verifiable justifications are critical for integrating automated tools into established review workflows.

Third, this study provided a detailed, step-by-step guide for developing and applying the tutorial, offering a practical framework that other teams can readily adopt. The guide outlines how to prepare inclusion and exclusion criteria for ingestion, conduct comprehension checks to ensure accurate interpretation, iteratively refine prompts with targeted feedback, and determine operational parameters. By documenting these procedures in a transparent and reproducible manner, the tutorial not only demonstrates feasibility but also serves as a transferable model that research groups with varying levels of technical expertise can implement in their own systematic reviews.

Fourth, our findings highlight the importance of operational parameters such as batch size and file format. For example, in our evaluation, limiting abstract screening batches to 50 articles in plain-text format optimized accuracy and minimized formatting-related errors. While this number reflects current agent performance, it should be regarded as a context-specific benchmark rather than a fixed rule, as future LLM developments may support larger inputs without compromising accuracy.

Finally, by using datasets from three meta-analyses that varied in topic domain, methodological design, eligibility criteria, and exclusion logic, this study showed that LLM maintained strong performance in article screening across contexts. These findings suggest that the workflow's effectiveness is not merely an artifact of domain-specific familiarity or the structural regularity of RCT reporting, but reflects a more generalizable capacity to apply nuanced, natural-language eligibility criteria across diverse forms of evidence.

Despite these strengths, several limitations warrant consideration. First, we evaluated only one model (ChatGPT-4o) within a specific interface and configuration; results may vary across agent versions, platforms, and future updates. Second, although we achieved high sensitivity and good specificity, the workflow still requires human oversight to verify inclusion/exclusion decisions and to catch rare but consequential misclassifications. The single false negative observed in Study 3 at the full-text stage serves as a reminder that LLM-assisted screening should be regarded as a decision-support tool rather than a replacement for human judgment. Third, while the iterative prompt refinement process improved performance, it also required initial human effort to identify and correct misinterpretations, suggesting that some investment in setup is necessary before efficiency gains can be realized. Fourth, while Study 3 extends the validation to observational, correlational evidence, the findings may have limited transferability to other specialized review types in healthcare, including reviews of diagnostic test accuracy, prognostic factors, clinical prediction models, and cost-effectiveness. These review types employ distinct evidence structures and reporting frameworks, such as QUADAS-2, PROBAST, and CHEERS, that differ substantially from the eligibility criteria structures examined across our three studies. To our knowledge, direct evidence on LLM-assisted screening in these contexts remains absent from the literature, and we encourage future research to evaluate performance across this broader range of review types.

Future research should continue to expand the evidence base across diverse review topics, outcome domains, and LLM platforms. Additionally, integrating LLM-assisted screening with established review management tools such as Covidence and Rayyan could further streamline workflows and facilitate hybrid human–AI screening pipelines. The potential for LLMs to assist with other review stages, including data extraction, risk-of-bias assessment, and evidence synthesis, also warrants exploration to assess their broader role in accelerating evidence-based research.

Conclusion

We developed a structured, prompt-based tutorial for article screening in systematic reviews and found that it achieved high sensitivity and good specificity at both abstract and full-text screening. This study offered a scalable, accessible, and adaptable approach to reduce workload of systematic reviews and meta-analyses. While human oversight remains essential, the integration of LLMs into screening workflows holds significant potential to reduce reviewer burden, maintain methodological rigor, and accelerate the evidence synthesis process.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0266462326104073.

Data availability statement. Data available on request from the authors.

Author contribution. T. Q.: conceptualization, formal analysis, methodology, project administration, visualization, writing – original draft, writing – review and editing. H. Q.: conceptualization, methodology, supervision, writing – review and editing. M. C: data curation, methodology, and writing – review and editing. T. L.: conceptualization, data curation, methodology, resources, supervision, writing – review and editing.

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

Competing interests. The authors declare that no competing interests exist.

References

1. **Gough D**, **Elbourne D**. Systematic research synthesis to inform policy, practice and democratic debate. *Soc Policy Soc*. 2002;**1**(3):225-236. https://doi.org/10.1017/S147474640200307X

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 09:23:08, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0266462326104073

2. **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.
3. **Qureshi R**, **Shaughnessy D**, **Gill KAR**, et al. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst Rev*. 2023;**12**(1):72. https://doi.org/10.1186/s13643-023-02243-z.
4. **Tsafnat G**, **Glasziou P**, **Choong MK**, et al. Systematic review automation technologies. *Syst Rev*. 2014;**3**(1):74. https://doi.org/10.1186/2046-4053-3-74
5. **Chai KEK**, **Lines RLJ**, **Gucciardi DF**, **Ng L**. Research screener: A machine learning tool to semi-automate abstract screening for systematic reviews. *Syst Rev*. 2021;**10**(1):93. https://doi.org/10.1186/s13643-021-01635-3
6. **Moher D**, **Liberati A**, **Tetzlaff J**, **Altman DG**, **Group TP**. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*. 2009;**6**(7):e1000097. https://doi.org/10.1371/journal.pmed.1000097
7. **Campos DG**, **Fütterer T**, **Gfrörer T**, et al. Screening smarter, not harder: A comparative analysis of machine learning screening algorithms and heuristic stopping criteria for systematic reviews in educational research. *Educ Psychol Rev*. 2024;**36**(1). https://doi.org/10.1007/s10648-024-09862-5
8. **Dennstädt F**, **Zink J**, **Putora PM**, **Hastings J**, **Cihoric N**. Title and abstract screening for literature reviews using large language models: An exploratory study in the biomedical domain. *Syst Rev*. 2024;**13**(1):158. https://doi.org/10.1186/s13643-024-02575-4
9. **Marshall IJ**, **Wallace BC**. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;**8**(1):163. https://doi.org/10.1186/s13643-019-1074-9
10. **O'Mara-Eves A**, **Thomas J**, **McNaught J**, **Miwa M**, **Ananiadou S**. Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Syst Rev*. 2015;**4**(1):5. https://doi.org/10.1186/2046-4053-4-5
11. **Galli C**, **Gavrilova AV**, **Calciolari E**. Large language models in systematic review screening: Opportunities, challenges, and methodological considerations. *Information*. 2025;**16**(5):378. https://doi.org/10.3390/info16050378
12. **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 [Internet] (EASE '24)*. New York, NY: Association for Computing Machinery; 2024 [cited 2025 Aug 12]. p. 262–271. https://doi.org/10.1145/3661167.3661172
13. **Shin J**, **Tang C**, **Mohati T**, et al. Prompt Engineering or Fine-Tuning: An Empirical Assessment of LLMs for Code [Internet]. arXiv; 2025 [cited 2025 Aug 12]. https://doi.org/10.48550/arXiv.2310.10508
14. **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
15. **Issaiy M**, **Ghanaati H**, **Kolahi S**, et al. Methodological insights into ChatGPT's screening performance in systematic reviews. *BMC Med Res Methodol*. 2024;**24**(1):78. https://doi.org/10.1186/s12874-024-02203-8
16. **Purewal A**, **Fautsch K**, **Klasova J**, **Hussain N**, **D'Souza RS**. Human versus artificial intelligence: Evaluating ChatGPT's performance in conducting published systematic reviews with meta-analysis in chronic pain research. *Reg Anesth Pain Med*. 2025. https://doi.org/10.1136/rapm-2024-106358.
17. **Cao C**, **Sang J**, **Arora R**, et al. Prompting is all you need: LLMs for systematic review screening [Internet]. medRxiv; 2024 [cited 2025 Aug 12]. p. 2024.06.01.24308323. https://doi.org/10.1101/2024.06.01.24308323
18. **Khraisha Q**, **Put S**, **Kappenberg J**, **Warraitch A**, **Hadfield K**. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res Synth Methods*. 2024;**15**(4):616-626. https://doi.org/10.1002/jrsm.1715
19. **Adhikari NS**, **Agarwal S.** A Comparative Study of PDF Parsing Tools Across Diverse Document Categories [Internet]. arXiv; 2025 [cited 2025 Sep 23]. https://doi.org/10.48550/arXiv.2410.09871
20. **Bast H**, **Korzen C.** A benchmark and evaluation for text extraction from PDF. In *2017 ACM/IEEE joint conference on digital libraries (JCDL) [Internet]* [cited 2025 Sep 23]. p. 1–10. https://doi.org/10.1109/JCDL.2017.7991564
21. **Lin T**, **Heckman TG**, **Anderson T**. The efficacy of synchronous teletherapy versus in-person therapy: A meta-analysis of randomized clinical trials. *Clin Psychol Sci Pract*. 2022;**29**(2):167-178. https://doi.org/10.1037/cps0000056
22. **Chen MS**, **Cai Q**, **Omari D**, et al. Emotion regulation and mental health across cultures: A systematic review and meta-analysis. *Nat Hum Behav*. 2025;**9**(6):1176-1200. https://doi.org/10.1038/s41562-025-02168-8
23. **Guo E**, **Gupta M**, **Deng J**, et al. Automated paper screening for clinical reviews using large language models: Data analysis study. *J Med Internet Res* 2024;**26**:e48996. https://doi.org/10.2196/48996.
24. **Li M**, **Sun J**, **Tan X**. Evaluating the effectiveness of large language models in abstract screening: A comparative analysis. *Syst Rev*. 2024;**13**(1):219. https://doi.org/10.1186/s13643-024-02609-x
25. **Baron S.** Trust, explainability and AI. *Philos Technol*. 2025;**38**(1):4. https://doi.org/10.1007/s13347-024-00837-6

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 09:23:08, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0266462326104073