Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Zero-To-Few Shot Large Language Model Performance in Systematic Review and Meta-Analysis Automation: A Validation Study Across All Major Steps Of Five Published Systematic Reviews

Mathias Pietrancosta

pietrancosta.mathias@gmail.com

Université de Montreal

Shishu Wang

Université de Montreal

Yichen Wu

Veronique Brulotte

Université de Montreal

Robert Harutyunyan

McGill University

Thomas Hemmerling

McGill University

Louis Morisson

Université de Montreal

Pascal Laferrière-Langlois

Université de Montreal

Research Article

Keywords: Large language models, Artificial intelligence, GPT-5, Systematic review automation, Evidence synthesis, Anesthesiology and Critical care.

Posted Date: October 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10879359/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Zero-To-Few Shot Large Language Model Performance in Systematic Review and Meta-Analysis Automation: A Validation Study Across All Major Steps Of Five Published Systematic Reviews.

Authors

Mathias Pietrancosta[1,2], Shishu Wang[1], Yichen Wu[3], Veronique Brulotte[1,2], Robert Harutyunyan[4], Thomas Hemmerling[4], Louis Morisson[1,2], Pascal Laferrière-Langlois[1,2].

Corresponding author

Mathias Pietrancosta, email: pietrancosta.mathias@gmail.com

Affiliations

[1]Faculty of Medicine, Université de Montreal, Montreal, QC, Canada.

[2]Department of Anesthesiology and Pain Medicine, Maisonneuve-Rosemont Hospital, Montreal, QC, Canada.

[3]Independent researcher.

[4]Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada.

Abstract

Background:

Systematic reviews and meta-analyses (SRMAs) represent the highest level of clinical evidence but remain highly labor-intensive. Large language models have emerged as tools to assist, and potentially automate, their production. Existing studies often rely on many-shot, task-specific prompting requiring expertise inaccessible to most researchers, limiting scalability. Conversely, fully agentic SRMA generation has proven unreliable, highlighting the need to identify performance bottlenecks requiring human intervention while automating the remaining tasks. This study evaluated the GPT-5 Pro series across all major

SRMA steps using a zero-to-few-shot pragmatic design to ensure reproducibility for non-AI specialists.

Methods:

Full data and metadata from five published and peer-reviewed SRMAs were extracted and processed. From October 2025 to February 2026, we prompted the GPT 5 Pro series through the web browser interface, to perform each major SRMA task, i.e. search string generation, article screening, data extraction, ROB assessment, and data synthesis, given the original author data from the previous step. For search string, article screening, and data synthesis steps, GPT outputs were compared to original author data. For data extraction and ROB assessment, GPT outputs were compared to a novel GPT-human hybrid reference standard.

Results:

GPT's search string sensitivity was 85% [80-88] $CI_{95\%}$. Abstract screening sensitivity and specificity were 80% [76-84] $CI_{95\%}$ and 85% [84-86] $CI_{95\%}$, respectively, while full-text screening underperformed at 46% sensitivity, [41-52] $CI_{95\%}$. Data extraction and ROB assessment accuracy were respectively 93% [90-95] $CI_{95\%}$ and 83% [79-86] $CI_{95\%}$. Meta-analytical result generation was meaningfully similar, though not strictly identical to original authors. All tasks were completed 10 to 100 times faster by GPT than by humans.

Discussion:

Zero-shot GPT 5 Pro outperforms humans in data extraction and ROB assessment. Conversely, for more "prompt-dependent" tasks, such as search string generation and article screening, GPT performs comparably, if not worse in the case of full-text screening, to humans. The latter tasks may not be accessible to true zero-shot prompting strategies, and require hybrid human-in-the-loop workflows for non-AI specialists. Finally, zero or few-shot GPT 5 Pro series data synthesis seems feasible, but hindered by technical considerations, which may impede its real-world usefulness.

Trial registration:

Not applicable.

Keywords:

Large language models, Artificial intelligence, GPT-5, Systematic review automation, Evidence synthesis, Anesthesiology and Critical care.

Background

Systematic reviews and meta-analyses (SRMAs) represent the highest bodies of evidence in the field of clinical research, corresponding to the systematic query, selection, appraisal, and synthesis of primary research results. SRMAs allow for the improvement of evidence-based practices as they consolidate research findings to inform decision-making and clinical guidelines(1). Despite the well-described and standardized workflow required to produce SRMAs, this literature is still both time-consuming and expensive, with estimates in the U.S.A. suggesting an average cost of $141,194.80 per review, urging the development of automation tools for this process(2). Large language models (LLMs), a form of natural language-based generative foundation models, have emerged in recent years as a potential tool for assisting(3), and potentially automatizing(4) the human labor involved in SRMA production. Because language is the primary medium of human knowledge and intelligence, we can leverage the ability of LLMs to generate human-like text-based outputs, given text-based inputs, into performing complex tasks, such as those involved in producing SRMAs.

The current literature focuses mostly on comparing human and LLM performance metrics for a single or a few key SRMA steps, especially article screening(5), data extraction(6) and risk of bias assessment(7). Many studies investigating the applications of LLMs in this context use them within an optimized framework, making use of prompt engineering, pre and post-processing data workflows, and sometimes integrating them into broader pipelines(4). More recently still, agentic LLM pipelines, capable of generating fully or quasi-fully automated SRMAs, have emerged as objects of study. While architectures which include human input at each step seem to exhibit strong performance(8–10), wholly unsupervised architectures are for the time being, unreliable(11,12). In this context, it becomes of chief concern to break down the process of SRMA generation into its major tasks, and identify performance bottlenecks which require human intervention, while fully automating the remaining tasks. Furthermore, while usage of such complex pipelines is necessary to pushing the boundaries of the artificial intelligence (AI) information landscape, this literature may not yield readily accessible changes in research practice to the uninitiated.

Hence, our approach is twofold. First, to provide a comprehensive overview of the GPT-5 Pro series' performance, the latest available model at the time of this study's inception, across all major steps of an SRMA: search query generation, article screening, data extraction, risk of bias assessment and data synthesis. Second, to provide the performance metrics of a zero-to-few shot, pragmatic use of GPT, in order to democratize the implementation of LLMs for non-AI specialists, using fully open-source prompts, data, and code.

Materials and methods

Overview

The objective of this present study is to evaluate the performance of the GPT-5 Pro series, for generating all major SRMA steps, given human-published data for the previous step. For our original data, we selected 5 published anesthesia-related SRMAs, all produced by the LIAM research center, for which raw data was fully available. These studies are hereafter referred to as the "Epidural SRMA"(13) on the topic of epidural hematomas in cardiac surgery with epidural catheterization, the "Gradient SRMA"(14), on the topic of central-to-peripheral arterial gradient risk factors in anesthesia and intensive care, the "LLM SRMA"(15), on the topic of large language model applications in anesthesia and intensive care, the "TIVA SRMA"(16), on the topic of safety and recovery profiles of total intravenous anesthesia versus inhalational anesthesia, and the "VR SRMA"(17), on the topic of virtual reality's impact on intraprocedural sedation and analgesia. We chose OpenAI's GPT-5 Pro series, the latest model available at the time of the study's inception, for its high performance against many scientific knowledge and intelligence-based benchmark sets, its accuracy across a large context-window, and near-unlimited prompting volume(18–20).

From October 2025 to February 2026, we accessed the GPT-5 Pro series through the ChatGPT web browser interface to generate each SRMA step. Default mode was toggled on for all modifiable GPT parameters, including usage of web-based tools. For each SRMA step, the latest available model was used; see Table 1.1. Inputs consisted of pre-drafted task-specific prompts, along with the corresponding clean task-specific data. Our general approach to prompt design was a zero-to-few shot strategy, requiring minimal optimization, using templates already available in the pertinent literature(21–23), to ensure reproducibility of usage and results for programming-naïve researchers. From a user's standpoint, replicating our GPT usage strategy requires no specific training, with human interventions mostly amounting to troubleshooting logical and format errors which require no programming knowledge. The comparator used as reference to measure GPT-5 Pro's performance against, for the search string, article screening and data synthesis steps,

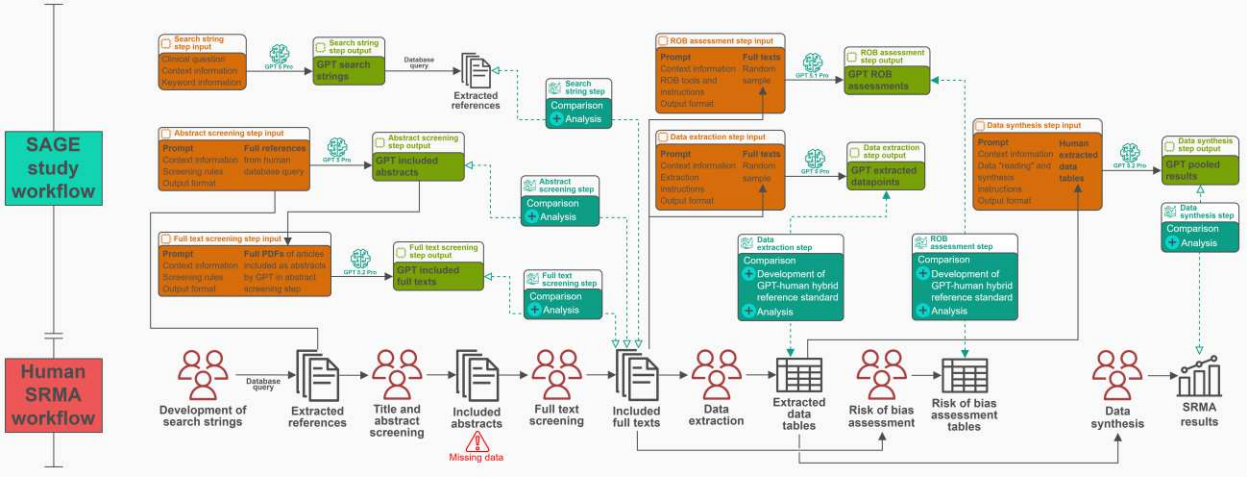
consisted of published original author data. Owing to risk of human error, when possible, blinded expert reviewers were recruited to produce more robust reference standards; see "Methods" sections "Data extraction step" and "Risk of bias assessment step". Table 1.1 showcases the inputs, outputs, comparators and metrics used to measure the GPT-5 Pro series' performance across all SRMA steps, as well as model use for each task. Figure 1.1 represents a simplified version of our study's workflow; full methodology and workflows are available in Supplementary Materials 1.1. All prompts are available in Supplementary Materials 1.2, while datasets and code are available at figshare(24).

Table 1.1: Overview of step-specific methodology for GPT-5 Pro series' inputs, outputs, comparators and performance metrics.

SRMA step	GPT input	GPT output	Comparison	Performance metrics
Search string generation	Prompt: clinical question for each SRMA and each database.	Search strings for each SRMA and each database.	All references retrieved from the database query using the GPT-5 Pro search string vs. all references retrieved from the original human database query. All references retrieved from the database query using the GPT-5 Pro search string vs. all references included in the final original human SRMA.	Sensitivity, precision and F1 score. SRMA-specific and pooled.
Article screening (abstracts)	Prompt: instructions for including and excluding abstracts, identical to original human instructions. Data: all references retrieved from original human database query.	GPT-5 Pro-included abstracts for each SRMA.	All abstracts included by GPT-5 Pro vs. all references included in the final original human SRMA.	Sensitivity, specificity, accuracy, precision, F1 score, Cohen's kappa and PABAK. SRMA-specific and pooled.
Article screening (full texts)	Prompt: instructions for including and excluding full-text articles, identical to original human instructions. Data: all abstracts included by GPT-5.2	GPT-5.2 Pro-included full text articles for each SRMA.	All full-text articles included by GPT-5.2 Pro vs. all references included in the final original human SRMA.	Sensitivity, specificity, accuracy, precision, F1 score, Cohen's kappa and PABAK. SRMA-specific and pooled.

	Pro in the previous abstract screening step.			
Data extraction	Prompt: instructions for data extraction, identical to original human instructions. Data: full text articles included in final original human SRMA.	GPT-5 Pro-generated data tables for each SRMA.	GPT-5 Pro-generated data tables vs. reference standard (human-GPT hybrid, see "Data extraction step").	Directionality of expert reviewer adjudication, accuracy, Cohen's kappa. SRMA-specific and pooled.
Risk of bias (ROB) assess-ment	Prompt: instructions for risk of bias assessment, including ROB tools. Data: full text articles included in final original human SRMA.	GPT-5.1 Pro-generated risk of bias assessments for each SRMA.	GPT-5.1 Pro-generated risk of bias assessments vs. reference standard (human-GPT hybrid, see "Risk of bias assessment step").	Directionality of expert reviewer adjudication, accuracy, Cohen's kappa, PABAK, F1 score, sensitivity and specificity. SRMA-specific and pooled.
Data synthesis	Prompt: instructions for calculating results. Data: human extracted data from full text articles included in final original human SRMA.	GPT-5.2 Pro-generated results.	GPT-5.2 Pro-generated results vs. results from the final original human SRMA.	Descriptive analysis.

Figure 1.1: Simplified study workflow for all SRMA steps combined.



The typical human research team's SRMA workflow is represented at the bottom, to the right of the red "Human SRMA workflow" panel. Above, to the right of the blue "SAGE study workflow" panel, is an abbreviated version of our study's workflow, across all SRMA steps. Orange boxes represent GPT inputs, green boxes represent GPT outputs, and blue boxes with corresponding dotted blue arrows represent comparison and analysis between GPT outputs and human data, detailed for every step in Table 1.1. Dark arrows represent processing of information, where the data contained in box at the beginning of the arrow is required to produce the data in the box at the end of the arrow.

Data acquisition

We included 5 SRMAs, spanning diverse anesthesia and intensive care topics, as well as a variety of methodologies and data types, allowing for evaluation of the GPT-5 Pro series' performance across a wide array of tasks. The full data required for GPT's generation of each SRMA step and comparison with reference standards can be subdivided into 3 categories. First, data was collected from the original SRMAs themselves, their study protocols, non-published original data tables, as well as any additional information requested upon correspondence. Secondly, search string and article selection steps required full access to all references exported to Rayyan platforms along with reviewer inclusion and exclusion decisions(25). Lastly, full-text screening required manual retrieval of full-text medical articles, as automated retrieval proved insufficient.

Search string step

The present step investigates the ability of GPT-5 Pro to generate database-specific search strings, given as input a single prompt. Following simple design guidelines(23), the prompt

contained contextual information of the task, the central thematic question each SRMA is attempting to answer, and the database-specific keyword format required. An individual search string was generated for each database every SRMA had originally queried: Medline, Embase, PubMed, Web of Science and CENTRAL. GPT's outputs were complete database- and SRMA-specific search strings, which were used to query each database, yielding medical article references as results. We authorized a minimal level of prompt troubleshooting for erroneous keywords and for search strings yielding amounts of references orders of magnitude outside what would be commonly expected of a human query. As such, criteria for accepting outputs were simply: a functional search string resulting in reference retrieval, and a total reference number ranging from 150 to 1500. The content of the troubleshooting prompts can be found in Supplementary Materials 1.2.

We performed two separate analyses, comparing references extracted using GPT's search strings with the full original human extracted references, as well as with the original humans' final included articles. While the first case allows direct comparison between both queries' raw results, only a median of 2.9% of all extracted references end up being included in final SRMAs(26). It follows that the primary performance metric should be whether GPT produces a search string that captures the references ultimately included in the final review, rather than one that is similar to the original human search string, regardless of it yielding the limited number of pertinent articles. The second comparison, between GPT's search string results and the articles included in the final human SRMAs, addresses this specifically. References extracted using human and GPT search strings were processed and matched to produce confusion matrices, absent of true negatives (which represent the articles not captured by both human and GPT search strings, i.e. the quasi-entirety of each database). SRMA-specific and pooled performance metrics including sensitivity, specificity, and F1 scores were calculated using a specific Python script.

Article screening step

The present step investigates the ability of the GPT-5 Pro series to screen medical articles according to specific decision rules, given abstracts and full texts. We subdivided this task into two distinct sub-steps, analogous to the human process: a first abstract-based screening, followed by a second full text-based screening. Prompt design in both cases incorporated contextual information of the task, inclusion criteria, exclusion criteria, and a decision rule, identical to those the original human reviewers used, as well as instructions for the output format(21). GPT-5 Pro and GPT-5.2 Pro were respectively used for abstract and full text-screening, as they were the latest available versions at the time.

For the abstract screening sub-step of each SRMA, GPT's input included the prompt, as well as the Rayyan data for all the references extracted from the original authors' database queries during the previous search string step. Rayyan platform data consisted of titles, abstracts and metadata (author, journal, year…) for each reference. GPT's output was a file containing each included article's reference, associated with its pertinent metadata. Information distinguishing author abstract and author full-text screening decisions was unavailable, and as such we failed to retrieve the full texts of abstracts included by the original authors at the end of their abstract-screening stage. Hence, for the full text-screening sub-step, GPT's input consisted of the prompt, as well as all available full text articles, in pdf format, included as abstracts by GPT during the previous abstract screening sub-step. Full text articles were mostly manually retrieved, as automated systems(27) proved insufficient. GPT's output was a list of the included full text articles, subsequently manually concatenated with the corresponding metadata. It follows that because GPT only screened full texts it had itself included as abstracts, the resulting full text screening sub-step performance can be in fact interpreted as the performance of a fully autonomous SRMA abstract and full-text screening.

While error rates are reported as substantial for database queries and article screening (28–30), development of "ground-truth" reference standards for these steps was beyond the scope of our study. Consequently, evaluation of GPT's performance was conducted by matching its outputs to the articles included in the final human SRMA. This was performed by use of Rayyan "keys", unique numerical identifiers for each article reference imported into the platform, which allow for error-free matching of both LLM-generated and original human datasets. Performance metrics were selected based on our best appreciation of the current literature(5,31), consisting of SRMA-specific and pooled accuracy, sensitivity, specificity, precision, F1 scores, Cohen's kappa and PABAK, calculated from confusion matrices with specific Python scripts.

Data extraction step

The present step investigates the ability of GPT-5 Pro to extract data from full text medical articles, given specific data extraction instructions. GPT's input consisted of a prompt containing task-specific contextual information, instructions identical to those humans used for datapoint extraction, a template output format, as well as a random sample of full text pdfs from each SRMA. GPT's outputs were raw files containing extracted data from each sample of articles. We calculated a required sample size of 384 datapoints, based on the standard formula for sample size calculation for estimating a single proportion with a

specified margin of error(32). Expected agreement between LLMs and humans was set at 0.90(6,33,34).

Human data extraction accuracy is reported in medical literature as ranging from 66 to 86%(35–40). Consequentially, and informed by the methodology of Cao et al.(4), we developed a novel human-GPT hybrid reference standard, to be used as "ground-truth" for measuring GPT's performance against. To this effect, we developed a Python script to match all concordant datapoints between the original authors and the GPT-extracted data tables, and retrieve only discordant datapoints. These discordant datapoints were adjudicated by two blinded expert reviewers (VB and SW), none having participated in the SRMAs included within the current study. Disagreements between reviewers were resolved by consensus, and the adjudicated datapoints were merged with the previous concordant datapoints, to form a novel data extraction standard.

We reported the directionality of our human expert reviewers' adjudication decisions, as a measure of GPT's performance when compared to the original SRMA authors. For all original author-to-GPT discordant datapoints, we therefore reported the proportion of adjudicated values correctly extracted by GPT, the proportion of adjudicated values correctly extracted by the original authors, and the proportion of adjudicated values incorrectly extracted by both GPT and original authors. Subsequently, we calculated GPT-to-reference-standard performance metrics, including SRMA-specific and pooled accuracies, Cohen's kappas and PABAKs with 95% confidence intervals.

Risk of bias assessment step

The present step investigates the ability of GPT-5.1 Pro to assess risk of bias (ROB) of medical articles, given standard ROB tools as instructions. GPT's input consisted of a prompt containing task-specific contextual information and instructions for ROB assessments, ROB assessment tools (ROB 2 and ROBINS-I) and corresponding output template formats, as well as a random sample of full text pdfs from each SRMA. Exception was made for the "Overall judgement" domains, where the inputs were specifically tailored prompts, along with ROB tools and GPT's concatenated ROB assessments for all previous domains for each article. Owing to lengthiness and implementation complexity of the ROB tools' decisional algorithms, we sequentially prompted GPT to separately assess risk of bias for each domain across all studies. GPT's outputs were raw ROB assessment data tables. We calculated a required sample size of 384 datapoints, based on the same formula as used for the data extraction step(32), albeit with differing parameters, notably, an 0.5 expected LLM-to-human agreement(7,41–43). While we recognize all risk of bias assessments may

not be truly independent, use of the design effect formula for clustered trials was not feasible due to lack of intraclass correlation coefficient data in the literature(44–46).

Human interrater reliability (IRR) of ROB assessments has been reported in as slight to fair, with kappa coefficients spanning 0.06 to 0.37(47–49), while proportion of agreement is shown to range between 57 and 81%(50). To this effect, we developed a novel human-GPT hybrid reference standard, to be used as reference for evaluating GPT's performance against, in the same manner as previously described in the "Data extraction step" section. Both expert human reviewers (RH and SW) had previous ROB assessment experience in multiple anesthesia-related SRMAs and performed blinded adjudications of human-to-GPT discordant assessments, resolving discrepancies through consensus, to produce the final ROB assessment standard.

We first compared GPT's performance to the original SRMA authors' ROB assessments, by reporting the directionality of our human expert reviewers' adjudication decisions in the same manner as previously described in the "Data extraction step". Subsequently, we calculated GPT-to-reference standard performance metrics, including SRMA-specific, domain-specific and pooled accuracy, Cohen's kappa and PABAK with 95% confidence intervals. We also reported sensitivity, specificity and F1 scores with 95% confidence intervals for dichotomous low versus non-low (i.e. medium and high) ROB assessment decisions. An alluvial diagram reporting differential classification of ROB assessments across all domains for all studies in each of the original authors', reference standard and GPT datasets was generated using Claude Sonnet 4.5(51).

Data synthesis step

The present step investigates the ability of GPT-5.2 Pro to generate meta-analytical results from raw study-level extracted data. GPT's input was a prompt containing step-specific contextual information as well as detailed instructions for "reading" extracted data and generating pooled outcomes(52), extracted data identical to the original author's data tables from the "Data extraction" step, and a template output format. GPT's outputs were review-level meta-analytical results for primary and secondary outcomes, in numerical or text-based form.

Data collection involved selection of outcomes for which full original author extracted data was available, i.e. the entirety of data necessary to be able to reproduce the exact same results, if the same calculations were applied. The outcomes which met these requirements were all primary outcome results across all SRMAs, as well as most secondary outcome results. We assessed GPT's performance by comparing its pooled results outputs to the

original human-published pooled results. Unfortunately, scarcity of literature for SRMA data synthesis provides little framework for performance analysis. As such, we did not calculate any performance metrics by use of statistical testing, opting instead for a descriptive analysis, similarly to a previous study on network meta-analysis automatization(53). In the present case, statistical testing would have been inappropriate, since the present step constitutes a set of single calculations from a dataset, without any sampling variability or any hypothesis testing about a sample, whose results are supposedly deterministic. Hence, sample size was not calculated for this step either.

Lastly, original authors' results from the "Epidural", "TIVA" and "VR" SRMAs were numerical, e.g. pooled incidences, ratios, standardized mean differences, heterogeneities etc., whose comparison with GPT outputs was descriptive, and performed directly without adjudication. Conversely, the "LLM SRMA" did not conduct a meta-analysis, focusing instead on text-based results, e.g. study summaries, categorization of study types and outcomes, identifying meta-themes and performance comparisons. Such interpretative text-based results required adjudication from a blinded expert reviewer (SW), assessing the meaningful equivalence or difference between original authors' and GPT's results. Assessments were reported in terms of directionality of adjudication decisions, alongside raw results.

Statistical analysis

All code for data processing and statistical analysis was written in Python 3.14 language(54), using the statistical software packages SciPy(55) and Scikit-learn(56), on Jupyter Notebooks(57) which were shared between authors with Google Colaboratory(58). Code was both manually written LLM-assisted(59–61). We chose Claude Sonnet 4.5(51) for its high performance in programming benchmark testing(62). All code was subsequently manually reviewed. The statistical analysis metrics used in our study are described in Supplementary Materials 1.1(63–71).

Results

Search string step

Rayyan data from 4 of the 5 SRMAs included within this study were extracted, excluding the "TIVA SRMA" for incomplete data. The amounts of references extracted from the database queries using both the original authors' and GPT's search strings are reported in Supplementary Materials 2.1. Across all databases and SRMAs, GPT-5 Pro ran for a total of

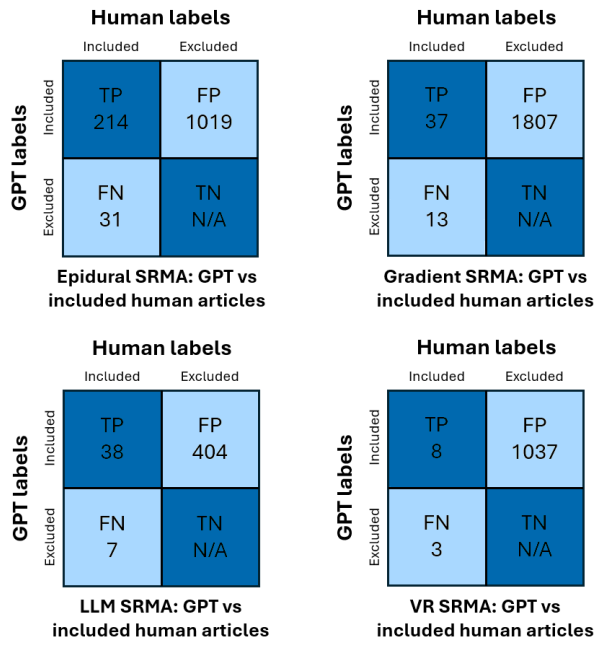
11 hours 3 min and 13 seconds. The average runtime for each search string generation was 33 minutes and 37 seconds, owing to an average of 2.3 prompts per output, mostly for sensitivity optimization and invalid keyword or syntax correction issues.

Table 2.1 reports GPT's search string performance metrics against the original human-included sets of references. Figure 2.1 represents the corresponding confusion matrices, excluding true negatives as explained in "Materials and methods". Pooled sensitivity for GPT 5 Pro search strings reach 0.85, 95% CI: 0.8 to 0.88. The other metrics such as precision and F1 scores are presented here as supplemental information. Performance metrics and confusion matrices for GPT-to-full author queries are available in Supplementary Materials 2.1.

Table 2.1: SRMA-specific and pooled performance metrics of database queries using GPT-5 Pro-generated search strings versus original human authors' included articles.

Dataset	Sensitivity [95% CI]	Precision [95% CI]	F1 score [95% CI]
Epidural SRMA	0.87 [0.83, 0.91]	0.17 [0.15, 0.2]	0.29 [0.26, 0.32]
Gradient SRMA	0.74 [0.6, 0.84]	0.02 [0.015, 0.028]	0.039 [0.027, 0.051]
LLM SRMA	0.84 [0.71, 0.92]	0.086 [0.063, 0.116]	0.16 [0.11, 0.2]
VR SRMA	0.73 [0.43, 0.9]	0.008 [0.004, 0.015]	0.015 [0.006, 0.026]
Pooled	0.85 [0.8, 0.88]	0.066 [0.058, 0.073]	0.12 [0.11, 0.13]

Figure 2.1: SRMA-specific confusion matrices for GPT search string results versus original authors' final included articles and versus full human search string results.



Article screening step

Rayyan data from 4 of the 5 SRMAs included within this study were extracted, excluding the "TIVA SRMA" for incomplete data. A total of 5845 references were screened during the abstract screening sub-step. For the subsequent full-text screening sub-step, all available full texts of the abstracts included by GPT during the abstract-screening step were manually retrieved, totaling 831. Figure 2.2 and 2.3 represent the abstract and full text screening confusion matrices for each SRMA. Full text retrieval flowchart and total screening results are available in Supplementary Materials 2.2.

Figure 2.2: SRMA-specific confusion matrices for GPT's included abstracts versus original authors' final included articles.

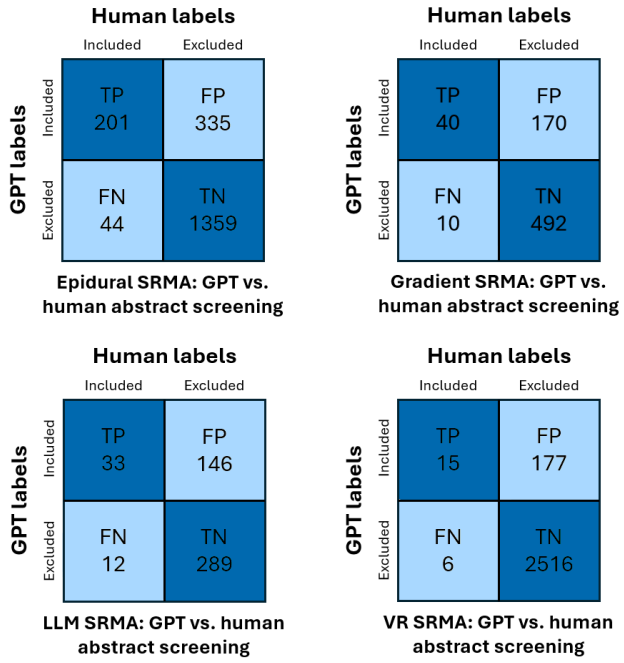
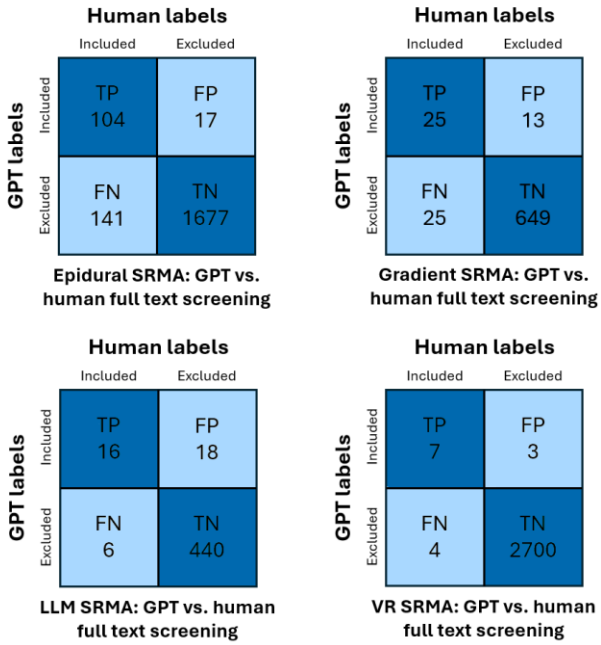


Figure 2.3: SRMA-specific confusion matrices for GPT's included full texts versus original authors' final included articles.



GPT runtime for abstract screening totaled 50 minutes and 9 seconds, averaging 0.51 seconds per screening decision. We report pooled sensitivity and specificity for GPT-5 Pro abstract screening to be respectively 0.8, 95% CI: 0.76 to 0.84, and 0.85, 95% CI: 0.84 to 0.86. Because abstract screening produces heavily imbalanced datasets, accuracy is best captured by pooled F1 score, only reaching 0.39, 95% CI: 0.36 to 0.42, due to its low precision. Interrater reliability between GPT and the original human reviewers is "fair", pooled Cohen's kappa 0.33, 95% CI: 0.3 to 0.36, and pooled PABAK 0.69, 95% CI: 0.67 to 0.71(67,68). Full abstract screening results are reported in Table 2.2.

Table 2.2: SRMA-specific and pooled performance metrics of GPT-5 Pro's abstract screening decisions versus original human authors' included articles.

Dataset	Sensitivity [95% CI]	Specificity [95% CI]	Accuracy [95% CI]	Precision [95% CI]	F1 score [95% CI]	Cohen's kappa [95% CI]	PABAK [95% CI]
Epidural SRMA	0.82 [0.77, 0.86]	0.8 [0.78, 0.82]	0.8 [0.79, 0.82]	0.38 [0.34, 0.42]	0.51 [0.47, 0.56]	0.41 [0.37, 0.46]	0.61 [0.57, 0.65]
Gradient SRMA	0.8 [0.67, 0.89]	0.74 [0.71, 0.78]	0.75 [0.71, 0.78]	0.19 [0.14, 0.25]	0.31 [0.23, 0.38]	0.22 [0.15, 0.28]	0.49 [0.43, 0.56]
LLM SRMA	0.73 [0.59, 0.84]	0.66 [0.62, 0.71]	0.67 [0.63, 0.71]	0.18 [0.13, 0.25]	0.29 [0.22, 0.37]	0.17 [0.1, 0.25]	0.34 [0.26, 0.43]
VR SRMA	0.71 [0.5, 0.86]	0.9 [0.92, 0.94]	0.93 [0.92, 0.94]	0.08 [0.05, 0.12]	0.14 [0.08, 0.21]	0.13 [0.07, 0.19]	0.87 [0.85, 0.88]
Pooled	0.8 [0.76, 0.84]	0.85 [0.84, 0.86]	0.85 [0.84, 0.86]	0.26 [0.23, 0.29]	0.39 [0.36, 0.42]	0.33 [0.3, 0.36]	0.69 [0.67, 0.71]

GPT runtime for full-text screening totaled 19 hours, 45 minutes and 33 seconds, averaging 1 minute and 28 seconds per screening decision. GPT-5.2 Pro's pooled sensitivity for full text screening, a composite of title, abstract and full-text screening, is low, at 0.46, 95% CI: 0.41 to 0.52, missing over half the pertinent articles. Conversely, pooled specificity reaches 0.99, 95% CI: 0.99 to 0.99. F1 score and precision are higher than for abstract screening, respectively pooled 0.57, 95% CI: 0.52 to 0.62, and 0.75, 95% CI: 0.68 to 0.8. Lastly, GPT-to-human pooled reviewer interrater reliability, both for Cohen's kappa and PABAK are respectively "moderate", 0.55, 95% CI: 0.5 to 0.6, and "almost perfect", 0.92, 95% CI: 0.91 to 0.93(67,68). Comprehensive full text screening results are reported in Table 2.3.

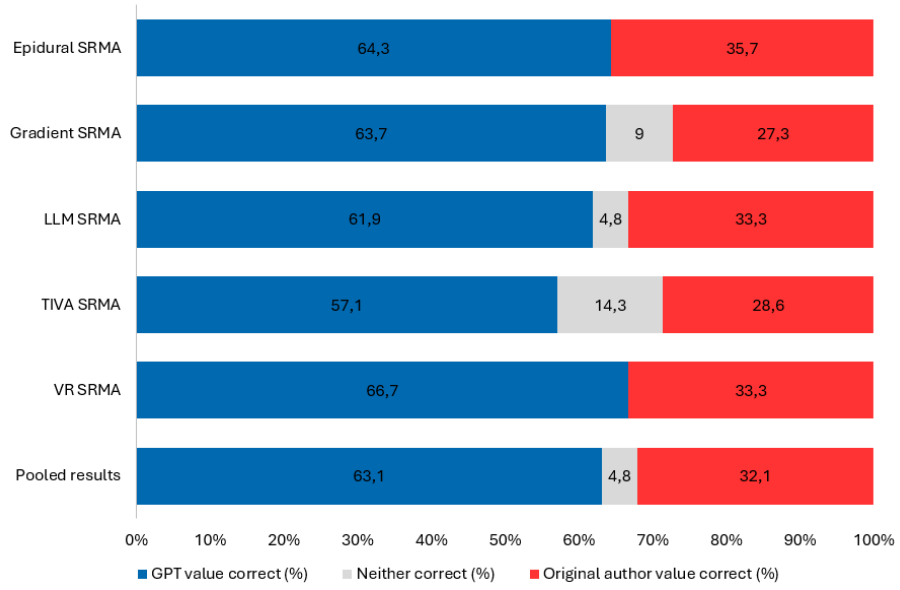
Table 2.3: SRMA-specific and pooled performance metrics of GPT-5.2 Pro's full text screening decisions versus original human authors' included articles.

Dataset	Sensitivity [95% CI]	Specificity [95% CI]	Accuracy [95% CI]	Precision [95% CI]	F1 score [95% CI]	Cohen's kappa [95% CI]	PABAK [95% CI]
Epidural SRMA	0.42 [0.36, 0.49]	0.99 [0.98, 0.99]	0.92 [0.91, 0.93]	0.86 [0.79, 0.91]	0.57 [0.51, 0.63]	0.53 [0.46, 0.59]	0.84 [0.81, 0.86]
Gradient SRMA	0.5 [0.37, 0.63]	0.98 [0.97, 0.99]	0.95 [0.93, 0.96]	0.66 [0.5, 0.79]	0.57 [0.43, 0.68]	0.54 [0.4, 0.66]	0.89 [0.86, 0.92]
LLM SRMA	0.73 [0.52, 0.87]	0.96 [0.94, 0.98]	0.95 [0.93, 0.97]	0.47 [0.31, 0.63]	0.57 [0.4, 0.71]	0.55 [0.37, 0.69]	0.9 [0.86, 0.94]
VR SRMA	0.64 [0.36, 0.85]	0.99 [0.99, 0.99]	0.99 [0.99, 0.99]	0.7 [0.4, 0.89]	0.67 [0.36, 0.88]	0.67 [0.36, 0.88]	0.99 [0.99, 0.99]
Pooled	0.46 [0.41, 0.52]	0.99 [0.99, 0.99]	0.96 [0.96, 0.97]	0.75 [0.68, 0.8]	0.57 [0.52, 0.62]	0.55 [0.5, 0.6]	0.92 [0.91, 0.93]

Data extraction step

Quality data from original authors' extracted datasets was available for all 5 SRMAs. Six studies from each SRMA were randomly selected, whose extracted data totaled 418 datapoints. Full datapoint characteristics are available in Supplementary Materials 2.3. Total GPT runtime was 1 hour and 33 minutes, corresponding to 3 minutes and 6 seconds per study, or 8.5 seconds per datapoint. Concordance matching analysis revealed a total of 84 discordant datapoints, amounting to a 12.8% discordance rate between GPT's and the original authors' extracted data tables. Each one of these datapoints was manually reviewed by two blinded expert reviewers (VB and SW), who in the final analysis agreed with GPT 63.1% of the time, while only agreeing with the original authors 32.1% of the time. Directionality of adjudication decisions are reported in Figure 2.4.

Figure 2.4: Lateral histogram representing direction of expert-reviewer data extraction adjudication process: agreement with GPT-5 Pro in blue, agreement with original human authors in red, agreement with neither in grey.



GPT-5 Pro's pooled datapoint extraction accuracy against the novel reference standard is 0.93, 95% CI: 0.9 to 0.95. GPT-to-reference standard data extraction interrater reliability is "almost perfect" as well, pooled 0.9, 95% CI: 0.82 to 0.96(67). PABAK is given here as supplemental information. Table 2.4 reports GPT's data extraction performance against the hybrid reference standard.

Table 2.4: SRMA-specific and pooled full text data extraction performance metrics of GPT-5 Pro versus the data extraction reference standard.

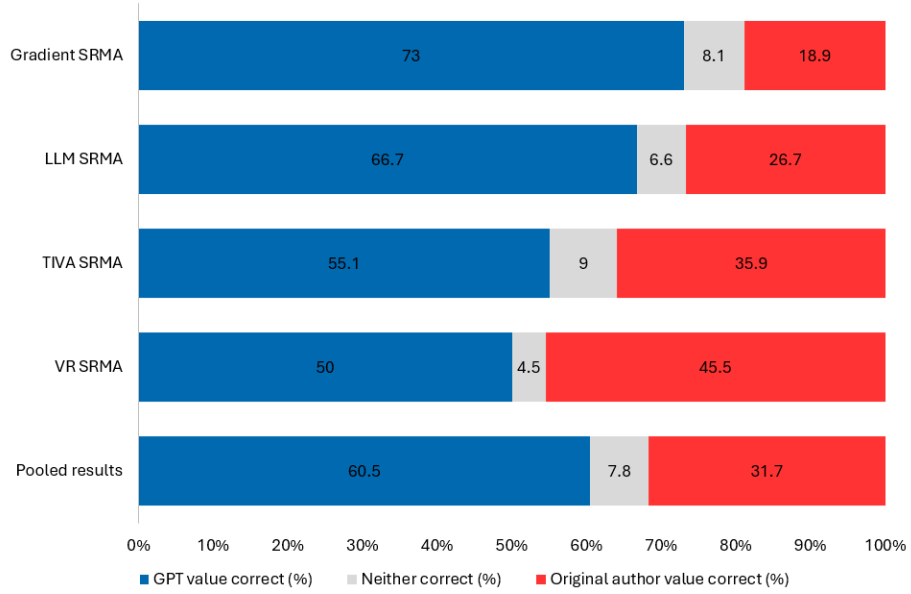
Dataset	Accuracy [95% CI]	Cohen's kappa [95% CI]	PABAK [95% CI]
Epidural SRMA	0.91 [0.8, 0.96]	0.87 [0.76, 0.96]	0.78 [0.59, 0.93]
Gradient SRMA	0.93 [0.82, 0.97]	0.92 [0.84, 0.98]	0.85 [0.7, 0.96]
LLM SRMA	0.91 [0.83, 0.95]	0.87 [0.8, 0.94]	0.76 [0.62, 0.89]
TIVA SRMA	0.9 [0.8, 0.95]	0.75 [0.64, 0.85]	0.53 [0.33, 0.73]
VR SRMA	0.95 [0.9, 0.97]	0.9 [0.82, 0.96]	0.9 [0.83, 0.96]
Pooled	0.93 [0.9, 0.95]	0.87 [0.81, 0.9]	0.85 [0.81, 0.88]

Risk of bias assessment step

Risk of bias assessments were available for 4 out of the 5 included SRMAs, from which 58 studies were randomly selected, totaling 398 individual risk of bias assessments. Two separate risk of bias tools were used by the original authors: ROB 2 for randomized clinical trials and ROBINS-I for non-randomized clinical trials. Full ROB data is available in Supplementary Materials 2.4. Total GPT runtime was 10 hours, 16 minutes and 4 seconds,

corresponding to an average of 10 minutes and 37 seconds per study. GPT runtime averaged 7 minutes and 5 seconds per study using the ROB 2 tool, and 15 minutes and 18 seconds per study using the ROBINS-I tool. Concordance matching analysis classified 42.8% of ROB assessments as discordant between GPT and original authors. Each of the 167 discordant assessments was manually reviewed by two blinded expert reviewers (RH and SW), classifying GPT as correct 60.5% of the time, and the original authors as correct only 31.7% of the time. One study was removed from the final reference standard(72), owing to incorrect use of ROB tools by the original authors. Directionality of adjudication decisions are reported in Figure 2.5.

Figure 2.5: Lateral histogram representing direction of expert-reviewer risk of bias assessments' adjudication process: agreement with GPT-5.1 Pro in blue, agreement with original human authors in red, agreement with neither in grey.

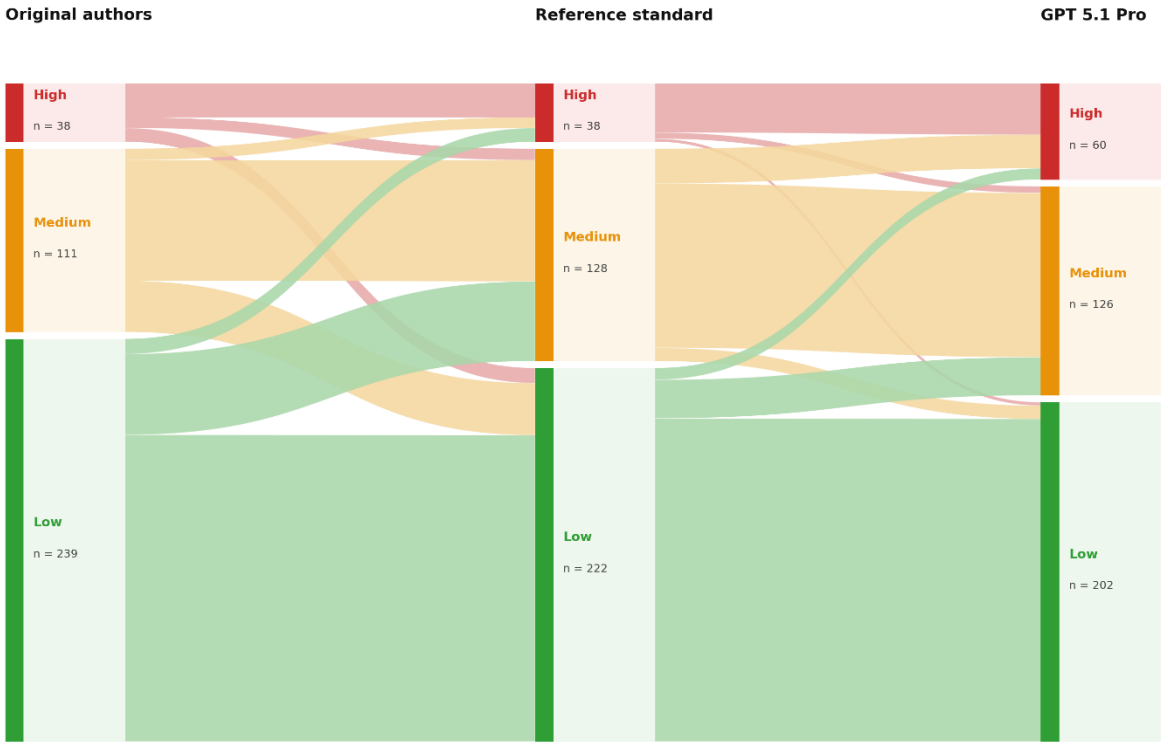


GPT-5.1 Pro's pooled accuracy against the novel reference standard is 0.83, 95% CI: 0.79 to 0.86. GPT-to-reference standard's interrater reliability is graded as "substantial", pooled Cohen's kappa at 0.66, 95% CI: 0.58 to 0.73(67). GPT's sensitivity, specificity and F1 scores for discriminating "low" versus "non-low" risk of bias assessments relative to reference standard are shown in Table 2.5. Furthermore, domain-specific accuracies for both ROB tools, pooled across all studies were calculated, and are fully available in Supplementary Materials 2.4. Lastly, an alluvial diagram was generated, aggregating all studies' individual ROB assessments across the "Original authors", "Reference standard" and "GPT 5.1 Pro" datasets. Figure 2.6 represents how the original authors, GPT and the reference standard assessed the same ROB domains for the same studies.

Table 2.5: SRMA-specific and pooled risk of bias assessment performance metrics of GPT-5.1 Pro versus the ROB assessment reference standard. Sensitivity, specificity, Cohen's kappa, PABAK and F1 scores reflect the ability of GPT to discriminate between "low" and "non-low" risk of bias assessments.

Dataset	Accuracy [95% CI]	Sensitivity [95% CI]	Specificity [95% CI]	F1 score [95% CI]	Cohen's kappa [95% CI]
Gradient SRMA	0.89 [0.8, 0.94]	0.91 [0.8, 0.96]	0.97 [0.86, 0.99]	0.94 [0.89, 0.98]	0.77 [0.64, 0.89]
LLM SRMA	0.89 [0.8, 0.94]	0.93 [0.84, 0.97]	0.85 [0.68, 0.94]	0.93 [0.88, 0.98]	0.77 [0.64, 0.89]
TIVA SRMA	0.79 [0.72, 0.85]	0.81 [0.71, 0.88]	0.96 [0.9, 0.99]	0.88 [0.82, 0.93]	0.58 [0.46, 0.7]
VR SRMA	0.76 [0.62, 0.86]	0.81 [0.6, 0.92]	0.92 [0.75, 0.98]	0.85 [0.71, 0.95]	0.52 [0.26, 0.74]
Pooled	0.83 [0.79, 0.86]	0.87 [0.81, 0.9]	0.94 [0.89, 0.97]	0.91 [0.88, 0.93]	0.66 [0.58, 0.73]

Figure 2.6: Alluvial diagram of individual risk of bias assessments for each study in the "Original authors", "Reference standard" and "GPT 5.1 Pro" datasets. For the purpose of this diagram, ROB assessments were systematized in the following manner: "Low" corresponding to ROBINS-I and ROB 2's "low" assessments, "Medium" corresponding to ROBINS-I and ROB 2's "some concerns" and "moderate" assessments, and "High" corresponding to ROBINS-I and ROB 2's "high", "serious" and "critical" assessments.



Data synthesis step

Full extracted data necessary to replicate human published results was available in 4 out of the 5 included SRMAs. Pooled GPT runtimes amounted to an average of 45.9 seconds per meta-analytic result. Characteristics of meta-analytical results and GPT runtime details are available in Supplementary Materials 2.5.1. GPT's numerical-based data synthesis results, including primary statistics, 95% confidence intervals and heterogeneity metrics, were meaningfully but not always strictly identical to those produced by the original authors manually inputting data into RevMan(73). Heterogeneity results, such as Chi^2 and degrees of freedom, reflected in the I^2 metric, were also meaningfully but not always strictly identical between humans and GPT. All sample size results data, including number of studies and participants in each group were strictly identical between original human authors and GPT (data not shown). Table 2.6 reports GPTs' and original authors' main data synthesis results. Data not shown is fully available in Supplementary Materials 2.5.2.

Table 2.6: Main data synthesis numerical-based results of original authors and GPT-5.2 Pro.

SRMA	Outcome	Human result	GPT-5.2 Pro result
Epidural	Hematoma incidence in RCTs	0, 95% CI [0–0.0014]	0, 95% CI [0–0.0014]

	Hematoma incidence in RCTs + prospective trials	0, 95% CI [0–0.00028]	0, 95% CI [0–0.00028]
	Hematoma incidence in RCTs + prospective + retrospective trials	0, 95% CI [0–0.000098]	0, 95% CI [0–0.000098]
	Hematoma incidence in RCTs + prospective + retrospective trials + case reports	0.00012, 95% CI [0.0000024–0.00024]	0.00012, 95% CI [0.0000024–0.00024]
TIVA	Grade 3-4 adverse events risk ratio, 95% CI and I^2	1.00, 95% CI [0.88–1.12], I^2 = 37%	1.00, 95% CI [0.88–1.12], I^2 = 36%
	PONV risk ratio, 95% CI and I^2	1.49, 95% CI [1.38–1.60], I^2 = 41%	1.48, 95% CI [1.37–1.59], I^2 = 39%
	Pain scores standardized mean difference, 95% CI and I^2	0.06, 95% CI [-0.03–0.16], I^2 = 62%	0.06, 95% CI [-0.03–0.16], I^2 = 63%
	Agitation risk ratio, 95% CI and I^2	1.50, 95% CI [1.15–1.96], I^2 = 56%	1.50, 95% CI [1.15–1.96], I^2 = 56%
	Shivering risk ratio, 95% CI and I^2	0.91, 95% CI [0.79–1.06], I^2 = 7%	0.91, 95% CI [0.79–1.06], I^2 = 7%
VR	Propofol consumption standardized mean difference, 95% CI and I^2	-1.7, 95% CI [-3.10– -0.31], I^2 = 92%	-1.71, 95% CI [-3.09– -0.33], I^2 = 92%
	Midazolam consumption standardized mean difference, 95% CI and I^2	-0.29, 95% CI [-0.57–0], I^2 = 18%	-0.29, 95% CI [-0.57– -0.01], I^2 = 20%
	Post-anesthesia care unit length of stay, 95% CI and I^2	-0.52, 95% CI [-0.86– -0.19], I^2 = 0%	-0.53, 95% CI [-0.86– -0.19], I^2 = 0%

Text-based results of the "LLM SRMA" were subdivided into 3 categories: answering of high-level questions, LLM-to-human performance comparisons and study summaries. Figure 2.7 represents the directionality of reviewer decisions for GPT and original author data synthesis results. In the final analysis, the expert found 40.8% of results to be meaningfully identical between GPT and original authors, 29.6% of results assessed correctly only by GPT, 22.2% of results assessed correctly only by original authors, and 7.4% of results assessed incorrectly by both GPT and original authors. Full GPT-5.2 Pro and original author data synthesis results are available in Supplementary Materials 2.5.2, while Table 2.7 reports examples for each of the categories of the "LLM SRMA"'s text-based results.

Figure 2.7: Lateral histogram representing direction of expert-reviewer data synthesis results adjudication process: equivalence between GPT and humans in purple, agreement with GPT in blue, agreement with original human authors in red, agreement with neither GPT nor humans in grey.

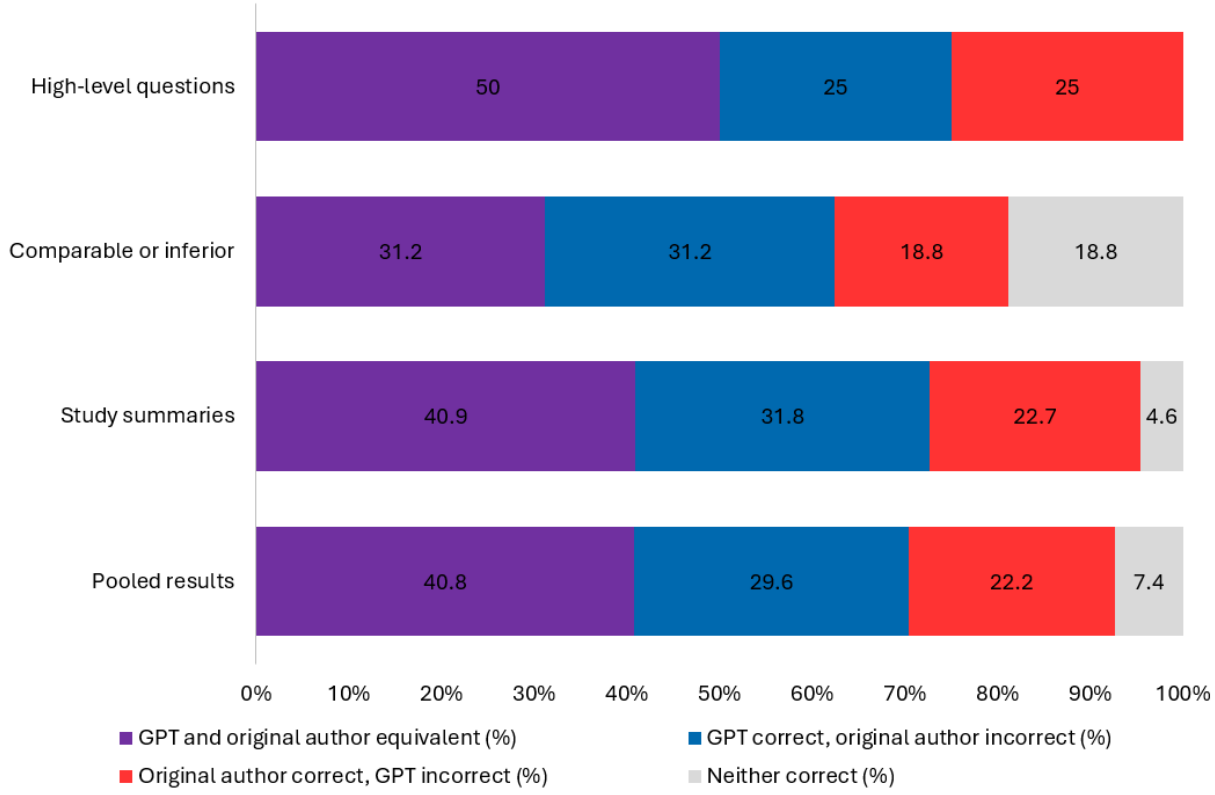


Table 2.7: Examples of data synthesis text-based results from original authors and GPT-5.2 Pro.

SRMA	Outcome	Human result	GPT-5.2 Pro result
LLM	Study summary for Akhondi-Asl 2024 reference	A fine-tuned version of LLaMA-7B showed a level of note quality superior to other models but inferior to the standard of physician.	LLM-generated PICU notes were rated worse than physician notes, with fine-tuned LLaMA performing best among the tested models.
	Can ChatGPT accurately predict clinical outcomes in patients in the ICU?	References: Liu 2024, Gakuba 2024, Chung 2024, Amacher 2024. GPT-4 showed variable performance, being comparable to physicians in intubation prediction, performing moderately in hospital admissions and mortality, and performing poorly in six month post-TBI outcome prediction and PACU phase 1 duration.	References: Amacher 2024, Gakuba 2024, Liu 2024. Across three ICU prognostication studies, GPT-4 achieved good discrimination for mortality/neurological outcome (AUROC 0.85) and performed similarly to established prognostic scores, and for HFNC patients its 48-h intubation-risk judgements were comparable to specialist physicians and better than non-specialists (GPT-4 > GPT-3.5). However, for 6-month

			TBI outcome prediction ChatGPT underperformed clinicians and the IMPACT model (AUROC 0.62 vs 0.70–0.71 and 0.81), indicating performance is outcome- and context-dependent.
	LLM-to-human comparison for clinical score prediction: number of studies	LLM comparable to humans: 3. LLM inferior to humans: 0.	LLM comparable to humans: 2. LLM inferior to humans: 3.

Discussion

GPT-to-human comparison

In this study, we report GPT-5 Pro as performing similarly to humans for search string generation, with reported sensitivities varying from 75% for conceptual approaches, to 97% for objective approaches, using Cochrane reviews as reference standards(74,75). Data for duration of human search string development is scarce, precluding comparative analyses with GPT runtimes. GPT-5 Pro's pooled abstract screening sensitivity and specificity were also roughly comparable to humans, while F1 scores, precision and interrater reliability were inferior to human reviewers(76–79). The subsequent full text screening step, interpreted as GPT-5.2 Pro's fully automated abstract and full text screening, missed over half the pertinent articles, far underperforming human reviewers. All other metrics performed below human standards as well, aside from specificity(67,68,76,79,80). In both abstract and full text screening cases, the F1 score, best measure of accuracy for imbalanced datasets, falls short in machine learning standards(80) as a consequence of GPT's relatively low precision. GPT's abstract and full text screening runtimes ranged from 10 to over 100 times faster than human reviewers(81–83).

GPT-5 Pro outperforms humans in data extraction accuracy and IRR(35–40), corroborated by blinded expert reviewer adjudication of discordant datapoints aligning with GPT a significant majority of the time. Notably, runtime was approximately 10 to over 20 times faster than for humans(84–86). GPT-5.1 Pro's risk of bias assessments were more accurate and reliable than humans(47–50,87–89), again corroborated by expert adjudication agreeing with GPT a significant majority of the time. GPT seems to generally match or outperform human accuracy in every risk of bias domain(50,87,88), except for ROB 2's overall risk of bias domain assessments, owing to differential interpretations of the tool; see

Supplementary Materials 2.4. Sensitivity, specificity and F1 scores for discriminating "low" versus "non-low" risk of bias assessments relative to the reference standard are remarkably high, albeit unavailable for comparison with human performances owing to scarcity of literature. Similarly to the data extraction step, runtime was approximately 10 times faster than for human reviewers(48,49,90–93). Lastly, GPT-5.2 Pro generated numerical meta-analytic results which were meaningfully but not strictly identical to the original authors' results, which may be compatible with numerous data handling issues and arithmetic effects; see Supplementary Materials 3.1. Text-based results showed a mixed profile of equivalence, superiority and inferiority between GPT and original authors. Pooled GPT runtimes averaged 45.9 seconds per meta-analytic result, which, although formal timelapse data concerning RevMan data synthesis from human authors is lacking, seems empirically much faster than what human authors are capable of.

Table 2.8: Head-to-head comparison of the GPT-5 Pro series and human performance metrics and runtimes, as reported in the literature, for each major SRMA step.

SRMA step	GPT-5 Pro series performance	Human reviewer performance	GPT-5 Pro series runtime	Human reviewer runtime
Search string generation	**Sensitivity** 85%	**Sensitivity** 75-97%(74,75)	**33 min 37 sec**/string	N/A
Abstract screening	**Sensitivity** 80% **Specificity** 85% **Precision** 26% **Cohen's kappa** 33%	**Sensitivity** 75-87%(76–78) **Specificity** 79-99%(76–78) **Precision** 91-98%(76,77) **Cohen's kappa** 65%(79)	**0.51 sec**/abstract **12 min 32 sec**/SRMA	**2 min**/abstract **335** abstracts/day **33 days**/SRMA(81–83)
Full text screening	**Sensitivity** 46% **Specificity** 99% **Precision** 75% **Cohen's kappa** 55%	**Sensitivity** 99%(76) **Specificity** 100%(76) **Precision** 99%(76) **Cohen's kappa** 75%(79)	**1 min 28 sec**/article	**20 min**/article(82)
Data extraction	**Accuracy** 93% **Cohen's kappa** 87%	**Accuracy** 66-86%(35–40) **Cohen's kappa** 88%(29)	**3 min 6 seconds**/article	**36-77 min**/article(84–86)
Risk of bias assessment	**Accuracy** 83% **Cohen's kappa** 66%	**Accuracy** 56-81%(50,87,88) **Cohen's kappa** 4-55%(47–49,89)	**7 min 5 sec**/article (ROB 2)	**2 hours**/article (ROB 2)(90–92) **28-180 min**/article (ROBINS-I)(48,93)

			15 min 18 sec/article (ROBINS-I)	
Data synthesis	For comparison, see Table 2.6 and Figure 2.7	For comparison, see Table 2.6 and Figure 2.7	**46 seconds**/result	N/A

Interpretation within the current literature

Different LLMs and implementation strategies markedly affect performance, ranging from poor(43) to quasi-fully automated SRMAs surpassing human reviewer capabilities(4). Some tasks seem less readily automatable, such as search string generation, owing to intrinsic limitations of LLMs, such as their lack of world models for "understanding" query concepts, and their failure to generate database-specific formalized query language. Furthermore, the LLM generation of search strings, much analogous to the human process of iterative development informed by feedback, may not be suited to a zero-shot approach. A more successful approach seems to be a multi-tiered prompt design, first generating keywords according to the clinical question, then organizing them into a search string, and lastly tuning for usable results(94). While more complex prompt-engineering solutions exist(95), even they seldom exceed 80% recall. For the time being, we believe GPT search string generation to be of limited usefulness, remaining helpful in producing the general framework and key terms of a Boolean query, but requiring human oversight to implement in practice. GPT's task-specific implementation challenges are fully discussed in Supplementary Materials 3.1.

Article screening may also be less readily automatable, owing to the task's "prompt-dependency", where performance is highly dependent on minor prompt modifications. While this study reports an acceptable performance of zero-to-few shot abstract screening, higher performance requires a level of prompt engineering beyond the resources of non-AI specialists(4,22,31,96,97). Hybrid human-LLM screening workflows have also been developed, using the LLM as a second reviewer for example, reportedly reducing workload by 33 to 93%(3), or as a summarizer of abstracts, reportedly increasing performance, and reducing workload by 75%(98). These emerging methods facilitate the human reviewing process while bypassing prompt engineering requirements. Supported by this study's results, it is our opinion that zero or few-shot GPT abstract screening is feasible and scalable, while fully automated article screening may not yet be as reproducible. Special attention must be paid to prompt development however, as the task's "prompt-dependency" prohibits a "single size fits all" approach, where a certain level of tailoring remains mandatory.

Study-level data extraction represents the most readily automatable task in our opinion, as performances both reported in the literature and in the present study reach or outclass human capabilities(4,6,33,34). This performance is further compounded by the reduction of human time and labor associated with its use(84–86). Data extraction can be reliably automated by non-AI specialists, given already existing prompt templates, such as those provided in our Supplementary Materials 1.2. To our knowledge, it seems a human-LLM hybrid might constitute an even more robust methodological design, with a first reviewer performing the task alongside the LLM, and a second, senior reviewer adjudicating only discordant datapoints. This design would ensure harnessing the high performance of LLMs, while minimizing the risk of residual errors.

Risk of bias assessment may also be just as readily automatable as data extraction, with performance metrics, both reported in our study and in the current literature, nearing or even outclassing human capabilities(7,41,42,91,99). Moreover, use of LLMs is associated with a drastic reduction of workload, performing the same task in approximately a tenth of human timeframes(48,49,90–93). Qualitatively, our findings (see Figure 2.6) also suggest that GPT's errors overwhelmingly indicate a pattern of systematic overestimation of bias, relative to the reference standard, unlike that of the original authors' assessments, which misclassify risk of bias in a non-systematic fashion. Analogous to our understanding of Type I errors being more dangerous than Type II errors in biomedical statistical analyses(100), falsely concluding that a "medium" or "high" risk of bias is actually "low", is of greater threat to scientific integrity than the opposite. In this instance, we feel more secure in using an artificial tool which systematically overestimates risk of bias in a small proportion of cases, than relying solely on human review, which may commit errors of non-predictable directionality. Here again, an even more robust methodology might be a human-LLM hybrid design, much akin to our data extraction step recommendations.

GPT's numerical-based meta-analytics exhibit acceptable performance in the present study, without being totally satisfactory, as there persists certain small differences between results generated by human use of RevMan and by GPT. These may be explained multifactorially by interpretation of ambiguous or missing data, as well as by intrinsic computational issues. While LLM data synthesis literature is scarce, our results are corroborated by Reason T et al.'s study on network meta-analyses, where GPT-4's results are nearly but not always strictly identical to humans'(53). Hence, while it is our opinion that numerical data synthesis could be fully automated with minimal risk to statistical significance, magnitude and directionality of results, very detailed prompting is required. Such prompts must meticulously define handling of missing and ambiguous data, as well as statistical model and sub-model use for each datapoint, perhaps rendering LLM use so burdensome as to hinder its labor reduction benefits.

Strengths of our study

We consider internal validity to be strong, as our study included 5 SRMAs, for which full original data and metadata were available for no less than 4 reviews for each step. Quality and comprehensiveness of original data was also remarkable, seldom requiring missing data imputations. The quantity of data required for each task, such as size of file attachments (maximum capacity of 10 full text articles per input), prohibited using a single input for each task. Instead, we broke down each task into batches of inputs, only differing by file attachments for example, a strategy which in effect, averaged GPT's performance across many iterations, and 5 SRMAs, for each step. In total, GPT extracted 418 datapoints and produced 398 ROB assessments (in excess of sample size requirements), screened 5845 abstracts and 831 full texts, and generated 194 meta-analytical results. Additionally, we minimized bias through use of Python scripts for concordance matching of datapoints and article references, and by prohibiting any of the original SRMA authors from contributing to this study, with the exception of YW, involved only in data processing. Furthermore, considering the poor data extraction and ROB assessment performance of human reviewers(35–40,47–50), novel hybrid reference standards were developed, allowing for measurement of GPT's performance against our best approximation of "ground-truth" standards.

We further deem external validity to be high, as included SRMAs spanned a wide array of anesthesia-related topics. As such, the range of systematic queries, the stringency and nature of screening decision rules, use of different ROB tools, as well as the types of extracted data and subsequent meta-analytical results, allow for extrapolation to many other SRMAs. To increase contextual understanding of our findings, GPT's performance was reported in comparison with its' corresponding human counterpart's performance, which required a review of the literature for each SRMA step. Our findings are further corroborated by previously published LLM performances. Relative to the current literature, we report: higher sensitivity for GPT's search strings, albeit literature is scarce and methodologically different(95); comparable abstract-screening results(3,5,31,77,79,101); underperformance of full-text screening results(79,102); comparable data extraction accuracy(6,33,34,79,84,103); comparable ROB assessment accuracy and IRR(7,41–43,91,99); and similar data synthesis results(53). Lastly, we designed our study along a zero-to-few shot, "user-friendly" approach, to ensure reproducibility of our results by non-AI specialists. We believe this approach to maximize generalizability of results, as well as to accelerate growth and adoption of human-labor sparing tools, such as the GPT-5 Pro series.

Limitations of our study

We recognize multiple limitations to our study, namely, issues related to access of the GPT-5 Pro series through its web-browser interface. While this design allows for greater generalizability of results to programming-naïve researchers, it unfortunately impedes ability to control for temperature, system prompts, and tool use, all of which may affect determinism of outputs(104). However, the widespread adoption of LLM browser interfaces in medical literature(105), corroborated by emerging literature suggestive of performance equivalence between use of application performance interfaces (APIs) and browser interfaces for systematic review tasks(106), has comforted us in this design decision. In order to further generalize our results, we also left the default mode toggled on for all modifiable ChatGPT parameters, including use of web-based tools. This carries a theoretical risk of output contamination by the open access data of our included SRMAs, for the full-text screening, risk of bias assessment and data synthesis steps. However, we believe the risk to be very limited, as chain-of-thought reasoning processes showed no signs of web search on inspection, data was not fully published for the tasks at risk, and we found no significant disparities among tasks performed on published versus unpublished data.

Another limitation of our study relates to the absence of gold standards and original author workflow data. First, we lacked “ground-truth” reference standards for search string and article screening steps, relying instead on original author screening data as the standard. This mainly affects the false positive rate, since there remains the possibility of pertinent references extracted or included by GPT, which could have been erroneously excluded by the original authors. While novel data extraction and ROB assessment reference standards were developed, and evidently superior to original human author data as comparators, these still cannot be reasonably considered as true gold-standards. Next, search string development methods were not recorded, which may have hindered our subsequent prompt designs(75,107). Finally, for the full text screening step, retrieving all full texts from all database query references proved to be too high in human labor cost, since automated article extraction was not effective enough to offset manual retrieval. Therefore, we could not perform GPT screening on all references from each database query, although such an endeavor would likely have been too labor intensive to be of reproducible interest for standard SRMA publishing teams.

Lastly, the strength of a zero-shot, “drag and drop” approach might misrepresent actual interactions with the ChatGPT interface, which are necessarily punctuated by numerous file loading issues, context window problems, erroneous output formats etc. The suitability of a blanket zero-to-few shot approach for all steps may also be questioned. While zero-shot prompting is readily applicable for data extraction, ROB assessment and data synthesis

tasks, it may prove maladapted for search string generation and article screening. In the former case especially, a few-shot strategy was inevitable; see Supplementary Materials 3.1. We also need address the costs involved in the use of GPT-5 Pro series, which amount to 200 USD/month, and may present a barrier to entry for research teams with limited resources. In addition, we emphasize the hidden time costs of GPT use, as reported runtimes only represent the amount of time the LLM takes to produce an output from an input. However, this seldom reflects the time required to acquire actual usable outputs from LLMs, which includes pre-processing of files, prompt development and troubleshooting, dealing with platform interruption errors, file drag and drop time etc.

Conclusion

Our study reports that zero-shot GPT-5 Pro series outperforms human authors in data extraction and risk of bias assessments. While the GPT-5 Pro series' search string generation and article screening performances remain roughly comparable, if not inferior to humans in the case of full-text screening, these steps are not accessible to true zero-shot prompting strategies, and require hybrid human-in-the-loop workflows for non-AI specialists. Finally, zero or few-shot GPT-5 Pro series data synthesis seems feasible, but hindered by small output errors and lengthy requirements for statistical model specification and handling of incomplete data, which may impede its real-world usefulness.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: In accordance with BMC Research Methods reporting standards, the authors have made available to the public all prompts, datasets and code used during the process of our study. Specifically, full prompts and prompt templates for each SRMA step can be found in Supplementary Materials 1.2. Full original SRMA author and GPT-5 Pro series output datasets, expert reviewer adjudication datasets, Python data processing, matching and statistical analysis code is available at https://doi.org/10.6084/m9.figshare.c.8518413.

Competing interests: The authors declare that they have no competing interests.

Funding: The authors declared no financial support was received for this work.

Authors' contributions (CRediT statements): Mathias Pietrancosta Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – Original draft, Writing – Review & Editing, Visualization, Supervision; **Shishu Wang** Investigation, Data curation; **Yichen Wu** Software, Investigation, Data curation; **Veronique Brulotte** Investigation, Data curation; **Robert Harutyunyan** Investigation, Data curation; **Thomas Hemmerling** Project administration; **Louis Morisson** Resources, Writing – Review & Editing, Project administration; **Pascal Laferrière-Langlois** Conceptualization, Resources, Writing – Review & Editing, Supervision, Project administration. All authors read and approved of the final manuscript.

Acknowledgements: We thank the authors Hakim Chala, Nicolas Daccache, and Joe Zako for providing complimentary SRMA data upon request. We gratefully acknowledge the CIUSSS EMTL (Centre Intégré Universitaire de Santé et de Services Sociaux de l'Est-de-l'Ile-de-Montréal) librarian team, Marie-Maxim Sévigny, Fatma Mohammed, Chahinez Hammoum and Julie Gaudreau, for their efforts in manually retrieving full-text medical articles. We used the generative AI tool Claude Sonnet 4.5(51), by Anthropic, chosen for its high performance in programming benchmark testing(62), to generate Python code involved in data processing and statistical analysis. We used the generative AI tool Claude Opus 4.8(108), by Anthropic, to generate Figure 2.6.

Bibliography

1. Cook DJ, Greengold NL, Ellrodt AG, Weingarten SR. The relation between systematic reviews and practice guidelines. Ann Intern Med. 1 août 1997;127(3):210-6. doi:10.7326/0003-4819-127-3-199708010-00006 PubMed PMID: 9245227.

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. déc 2019;16:100443. doi:10.1016/j.conctc.2019.100443 PubMed PMID: 31497675; PubMed Central PMCID: PMC6722281.

3. Delgado-Chaves FM, Jennings MJ, Atalaia A, Wolff J, Horvath R, Mamdouh ZM, et al. Transforming literature screening: The emerging role of large language models in systematic reviews. Proc Natl Acad Sci U S A. 14 janv 2025;122(2):e2411962122. doi:10.1073/pnas.2411962122 PubMed PMID: 39761403; PubMed Central PMCID: PMC11745399.

4. Cao C, Arora R, Cento P, Manta K, Farahani E, Cecere M, et al. Automation of Systematic Reviews with Large Language Models [Internet]. medRxiv; 2025 [cité 10 mars 2026]. p. 2025.06.13.25329541. Disponible sur:

https://www.medrxiv.org/content/10.1101/2025.06.13.25329541v1 doi:10.1101/2025.06.13.25329541

5. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. J Med Internet Res. 12 janv 2024;26:e48996. doi:10.2196/48996 PubMed PMID: 38214966; PubMed Central PMCID: PMC10818236.

6. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods. 2024;15(4):616-26. doi:10.1002/jrsm.1715

7. Eisele-Metzger A, Lieberum JL, Toews M, Siemens W, Heilmeyer F, Haverkamp C, et al. Exploring the potential of Claude 2 for risk of bias assessment: Using a large language model to assess randomized controlled trials with RoB 2. Res Synth Methods. mai 2025;16(3):491-508. doi:10.1017/rsm.2025.12

8. Parkinson RH, Cerbone H, Mieskolainen M, Cao S, Wilson AD, Albacete S, et al. MetaBeeAI: an AI pipeline for structured evidence extraction from biological literature [Internet]. bioRxiv; 2026 [cité 19 juin 2026]. p. 2025.11.24.690154. Disponible sur: https://www.biorxiv.org/content/10.1101/2025.11.24.690154v2 doi:10.1101/2025.11.24.690154

9. Lee K, Paek H, Ofoegbu N, Rube S, Higashi MK, Dawoud D, et al. A4SLR: An Agentic Artificial Intelligence-Assisted Systematic Literature Review Framework to Augment Evidence Synthesis for Health Economics and Outcomes Research and Health Technology Assessment. Value Health. 1 nov 2025;28(11):1655-64. doi:10.1016/j.jval.2025.08.002

10. Wang Z, Cao L, Danek B, Jin Q, Lu Z, Sun J. Accelerating Clinical Evidence Synthesis with Large Language Models [Internet]. arXiv; 2024 [cité 19 juin 2026]. Disponible sur: http://arxiv.org/abs/2406.17755 doi:10.48550/arXiv.2406.17755

11. Padarha S, Kearns RO, Naidoo T, Yang L, Borchmann Ł, BŁaszczyk P, et al. Evaluating AI-based Scientific Knowledge Synthesis with Epidemiological Systematic Reviews [Internet]. arXiv; 2026 [cité 19 juin 2026]. Disponible sur: http://arxiv.org/abs/2603.22327 doi:10.48550/arXiv.2603.22327

12. Loke WT, Srinivasan S, Yang GD, Zou K, Ong AY, Zhu LZ, et al. Can 'Deep Research' agents and general AI agentic systems autonomously perform systematic review and meta-analysis? Eye. janv 2026;40(1):147-9. doi:10.1038/s41433-025-04138-w

13. Laferrière-Langlois P, Jeffries S, Harutyunyan R, Hemmerling TM. Epidural Catheterization in Cardiac Surgery: A Systematic Review and Risk Assessment of

Epidural Hematoma. Ann Card Anaesth. 2024;27(2):111-20. doi:10.4103/aca.aca_160_23 PubMed PMID: 38607874; PubMed Central PMCID: PMC11095789.

14. Chala H, Daccache N, Morisson L, Denault A, Laferrière-Langlois P. Risk factors for a central-to-peripheral arterial blood pressure gradient in adults: a systematic review. Can J Anesth Can Anesth. 1 nov 2025;72(11):1634-49. doi:10.1007/s12630-025-03058-3

15. Daccache N, Zako J, Morisson L, Laferrière-Langlois P. The applications of ChatGPT and other large language models in anesthesiology and critical care: a systematic review. Can J Anesth Can Anesth. 1 juin 2025;72(6):904-22. doi:10.1007/s12630-025-02973-9

16. Daccache N, Wu Y, Jeffries SD, Zako J, Harutyunyan R, Pelletier ED, et al. Safety and recovery profile of patients after inhalational anaesthesia versus target-controlled or manual total intravenous anaesthesia: a systematic review and meta-analysis of randomised controlled trials. Br J Anaesth. 1 mai 2025;134(5):1474-85. doi:10.1016/j.bja.2025.02.007 PubMed PMID: 40074622.

17. Zako J, Daccache N, Morisson L, Richebé P, Laferrière-Langlois P. Virtual reality for reduction of intraprocedural pharmacological sedation and analgesia in adult patients: A systematic review and meta-analysis. Anaesth Crit Care Pain Med. 1 avr 2025;44(2):101483. doi:10.1016/j.accpm.2025.101483

18. Stats LLM. LLM Stats [Internet]. 2025 [cité 21 janv 2026]. GPT-5.2: Pricing, Context Window, Benchmarks, and More. Disponible sur: https://llm-stats.com/models/gpt-5.2-2025-12-11

19. LLM Leaderboard - Comparison of over 100 AI models from OpenAI, Google, DeepSeek & others [Internet]. [cité 21 janv 2026]. Disponible sur: https://artificialanalysis.ai/leaderboards/models

20. Présentation de GPT-5.2 [Internet]. 2026 [cité 21 janv 2026]. Disponible sur: https://openai.com/fr-FR/index/introducing-gpt-5-2/

21. Galli C, Gavrilova AV, Calciolari E. Large Language Models in Systematic Review Screening: Opportunities, Challenges, and Methodological Considerations. Information. mai 2025;16(5):378. doi:10.3390/info16050378

22. 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. 15 juin 2024;13(1):158. doi:10.1186/s13643-024-02575-4

23. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 4 oct 2023;25:e50638. doi:10.2196/50638 PubMed PMID: 37792434; PubMed Central PMCID: PMC10585440.

24. Pietrancosta M. Systematic assessments of GPT 5 versus expert reviewers (SAGE) study datasets and code [Internet]. 4 juin 2026. doi:10.6084/m9.figshare.c.8518413.v1

25. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 5 déc 2016;5(1):210. doi:10.1186/s13643-016-0384-4

26. Sampson M, Tetzlaff J, Urquhart C. Precision of healthcare systematic review searches in a cross-sectional sample. Res Synth Methods. 2011;2(2):119-25. doi:10.1002/jrsm.42

27. The EndNote Team,. EndNote. Philadelphia, PA: Clarivate; 2013.

28. Salvador-Oliván JA, Marco-Cuenca G, Arquero-Avilés R. Errors in search strategies used in systematic reviews and their effects on information retrieval. J Med Libr Assoc JMLA. avr 2019;107(2):210-21. doi:10.5195/jmla.2019.567 PubMed PMID: 31019390; PubMed Central PMCID: PMC6466507.

29. Hanegraaf P, Wondimu A, Mosselman JJ, Jong R de, Abogunrin S, Queiros L, et al. Inter-reviewer reliability of human literature reviewing and implications for the introduction of machine-assisted systematic reviews: a mixed-methods review [Internet]. 1 mars 2024. doi:10.1136/bmjopen-2023-076912

30. Belur J, Tompson L, Thornton A, Simon M. Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making. Sociol Methods Res. 1 mai 2021;50(2):837-65. doi:10.1177/0049124118799372

31. Syriani E, David I, Kumar G. Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews. J Comput Lang. août 2024;80:101287. doi:10.1016/j.cola.2024.101287

32. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform. avr 2014;48:193-204. doi:10.1016/j.jbi.2014.02.013 PubMed PMID: 24582925.

33. Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, Crotty K, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. Res Synth Methods. juill 2024;15(4):576-89. doi:10.1002/jrsm.1710 PubMed PMID: 38432227.

34. Khan MA, Ayub U, Naqvi SAA, Khakwani KZR, Sipra Z bin R, Raina A, et al. Collaborative Large Language Models for Automated Data Extraction in Living Systematic Reviews. medRxiv. 23 sept 2024;2024.09.20.24314108. doi:10.1101/2024.09.20.24314108 PubMed PMID: 39399004; PubMed Central PMCID: PMC11469465.

35. Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. J Clin Epidemiol. mars 2010;63(3):289-98. doi:10.1016/j.jclinepi.2009.04.007 PubMed PMID: 19683413.

36. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol. juill 2006;59(7):697-703. doi:10.1016/j.jclinepi.2005.11.010 PubMed PMID: 16765272.

37. Carroll C, Scope A, Kaltenthaler E. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. BMC Res Notes. 17 déc 2013;6:539. doi:10.1186/1756-0500-6-539 PubMed PMID: 24344873; PubMed Central PMCID: PMC3878552.

38. Tang L, Wang R, Doi SAR, Furuya-Kanamori L, Lin L, Qin Z, et al. Double data extraction was insufficient for minimizing errors in evidence synthesis: a randomized controlled trial [Internet]. medRxiv; 2024 [cité 10 mars 2026]. p. 2023.10.16.23297056. Disponible sur: https://www.medrxiv.org/content/10.1101/2023.10.16.23297056v2 doi:10.1101/2023.10.16.23297056

39. Xu C, Yu T, Furuya-Kanamori L, Lin L, Zorzela L, Zhou X, et al. Validity of data extraction in evidence synthesis practice of adverse events: reproducibility study. BMJ. 10 mai 2022;377:e069155. doi:10.1136/bmj-2021-069155 PubMed PMID: 35537752; PubMed Central PMCID: PMC9086856.

40. Li T, Saldanha IJ, Jap J, Smith BT, Canner J, Hutfless SM, et al. A randomized trial provided new evidence on the accuracy and efficiency of traditional vs. electronically annotated abstraction approaches in systematic reviews. J Clin Epidemiol. nov 2019;115:77-89. doi:10.1016/j.jclinepi.2019.07.005 PubMed PMID: 31302205.

41. Lai H, Ge L, Sun M, Pan B, Huang J, Hou L, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA Netw Open. 1 mai 2024;7(5):e2412687. doi:10.1001/jamanetworkopen.2024.12687 PubMed PMID: 38776081; PubMed Central PMCID: PMC11112444.

42. Lai H, Liu J, Bai C, Liu H, Pan B, Luo X, et al. Language models for data extraction and risk of bias assessment in complementary medicine. Npj Digit Med. 31 janv 2025;8(1):74. doi:10.1038/s41746-025-01457-w

43. Hasan B, Saadi S, Rajjoub NS, Hegazi M, Al-Kordi M, Fleti F, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. BMJ Evid-Based Med. 22 nov 2024;29(6):394-8. doi:10.1136/bmjebm-2023-112597 PubMed PMID: 38383136.

44. Kish L. Survey sampling [Internet]. New York, J. Wiley; 1965 [cité 10 mars 2026]. 674 p. Disponible sur: http://archive.org/details/surveysampling00kish

45. Design Effect Components [Internet]. [cité 27 oct 2025]. Disponible sur: https://cran.r-project.org/web/packages/PracTools/vignettes/Design-effects.html

46. Hayes RJ, Moulton LH. Cluster Randomised Trials. 2e éd. New York: Chapman and Hall/CRC; 2017. 424 p. doi:10.4324/9781315370286

47. Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol. sept 2013;66(9):973-81. doi:10.1016/j.jclinepi.2012.07.005 PubMed PMID: 22981249.

48. Minozzi S, Cinquini M, Gianola S, Castellini G, Gerardi C, Banzi R. Risk of bias in nonrandomized studies of interventions showed low inter-rater reliability and challenges in its application. J Clin Epidemiol. août 2019;112:28-35. doi:10.1016/j.jclinepi.2019.04.001 PubMed PMID: 30981833.

49. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol. oct 2020;126:37-44. doi:10.1016/j.jclinepi.2020.06.015 PubMed PMID: 32562833.

50. Bertizzolo L, Bossuyt P, Atal I, Ravaud P, Dechartres A. Disagreements in risk of bias assessment for randomised controlled trials included in more than one Cochrane systematic reviews: a research on research study using cross-sectional design. BMJ Open. 1 avr 2019;9(4):e028382. doi:10.1136/bmjopen-2018-028382 PubMed PMID: 30940766; PubMed Central PMCID: PMC6500379.

51. Claude 4.5 Sonnet [Large language model] [Internet]. Anthropic; 2024. Disponible sur: https://www.anthropic.com/claude

52. Deeks JJ, Higgins JP. Statistical algorithms in Review Manager 5.

53. Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential

Application of Large Language Models. PharmacoEconomics Open. 10 févr 2024;8(2):205-20. doi:10.1007/s41669-024-00476-9 PubMed PMID: 38340277; PubMed Central PMCID: PMC10884375.

54. Python Software Foundation. Python [Internet]. 2026. ([Computer software]). Disponible sur: https://www.python.org/

55. Virtanen P, Gommers R, Oliphant T, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. 2020. 17(3):261-72. doi:10.1038/s41592-019-0686-2

56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(85):2825-30.

57. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. In: IOS Press [Internet]. 2016 [cité 10 mars 2026]. p. 87-90. Disponible sur: https://ui.adsabs.harvard.edu/abs/2016ppap.book...87K doi:10.3233/978-1-61499-649-1-87

58. Google. Google Colaboratory [Internet]. Disponible sur: https://colab.research.google.com/

59. Uspenski A. 42% of Code Is Now AI-Assisted! ShiftMag [Internet]. 12 févr 2026 [cité 12 mars 2026]. Disponible sur: https://shiftmag.dev/state-of-code-2025-7978/

60. AI Coding Assistant Statistics & Trends [2025] | Second Talent [Internet]. 16 oct 2025 [cité 12 mars 2026]. Disponible sur: https://www.secondtalent.com/resources/ai-coding-assistant-statistics/

61. EliteBrains matches elite developers with tech companies. [Internet]. [cité 12 mars 2026]. Disponible sur: https://www.elitebrains.com/

62. SWE-bench Leaderboards [Internet]. [cité 12 mars 2026]. Disponible sur: https://www.swebench.com/

63. GeeksforGeeks [Internet]. 17:10:00+00:00 [cité 30 janv 2026]. What is Specificity. Disponible sur: https://www.geeksforgeeks.org/machine-learning/what-is-specificity/

64. Baratloo A, Hosseini M, Negida A, El Ashal G. Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. Emergency. 2015;3(2):48-9. PubMed PMID: 26495380; PubMed Central PMCID: PMC4614595.

65. Google for Developers [Internet]. [cité 30 janv 2026]. Classification: Accuracy, recall, precision, and related metrics | Machine Learning. Disponible sur:

https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

66. F1 Score in Machine Learning: Intro & Calculation [Internet]. [cité 30 janv 2026]. Disponible sur: https://www.v7labs.com/blog/f1-score-guide

67. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 15 oct 2012;22(3):276-82. PubMed PMID: 23092060; PubMed Central PMCID: PMC3900052.

68. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1 mai 1993;46(5):423-9. doi:10.1016/0895-4356(93)90018-V PubMed PMID: 8501467.

69. Chen G, Faris P, Hemmelgarn B, Walker RL, Quan H. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. BMC Med Res Methodol. 21 janv 2009;9:5. doi:10.1186/1471-2288-9-5 PubMed PMID: 19159474; PubMed Central PMCID: PMC2636838.

70. Dunnigan K, Consulting S. Confidence Interval Calculation for Binomial Proportions.

71. Bootstrap Confidence Intervals [Internet]. [cité 31 janv 2026]. Disponible sur: https://acclab.github.io/bootstrap-confidence-intervals.html

72. Yahagi M, Hiruta R, Miyauchi C, Tanaka S, Taguchi A, Yaguchi Y. Comparison of Conventional Anesthesia Nurse Education and an Artificial Intelligence Chatbot (ChatGPT) Intervention on Preoperative Anxiety: A Randomized Controlled Trial. J Perianesthesia Nurs Off J Am Soc PeriAnesthesia Nurses. oct 2024;39(5):767-71. doi:10.1016/j.jopan.2023.12.005 PubMed PMID: 38520470.

73. Review Manager (RevMan) [Computer program]. Version 5.4. The Cochrane Collaboration, 2020.

74. Hausner E, Guddat C, Hermanns T, Lampert U, Waffenschmidt S. Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. J Clin Epidemiol. févr 2015;68(2):191-9. doi:10.1016/j.jclinepi.2014.09.016 PubMed PMID: 25464826.

75. Hausner E, Guddat C, Hermanns T, Lampert U, Waffenschmidt S. Prospective comparison of search strategies for systematic reviews: an objective approach yielded higher sensitivity than a conceptual one. J Clin Epidemiol. 1 sept 2016;77:118-24. doi:10.1016/j.jclinepi.2016.05.002

76. Wilson E, Cruz F, Maclean D, Ghanawi J, McCann SK, Brennan PM, et al. Screening for in vitro systematic reviews: a comparison of screening methods and training of a machine learning classifier. Clin Sci. 27 janv 2023;137(2):181-93. doi:10.1042/CS20220594 PubMed PMID: 36630537.

77. Sanghera R, Thirunavukarasu AJ, El Khoury M, O'Logbon J, Chen Y, Watt A, et al. High-performance automated abstract screening with large language model ensembles. J Am Med Inform Assoc JAMIA. 1 mai 2025;32(5):893-904. doi:10.1093/jamia/ocaf050 PubMed PMID: 40119675; PubMed Central PMCID: PMC12012331.

78. Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol. mai 2020;121:20-8. doi:10.1016/j.jclinepi.2020.01.005 PubMed PMID: 31972274.

79. Cassell K, Ologunowa A, Rastegar-Mojarad M, Chun B, Huang YL, Wang D, et al. Analysis of article screening and data extraction performance by an AI systematic literature review platform. Front Artif Intell. 2025;8:1662202. doi:10.3389/frai.2025.1662202 PubMed PMID: 41356671; PubMed Central PMCID: PMC12676488.

80. F1 Score in Machine Learning [Internet]. [cité 10 mars 2026]. Disponible sur: https://encord.com/blog/f1-score-in-machine-learning/

81. Haddaway NR, Westgate MJ. Predicting the time needed for environmental systematic reviews and systematic maps. Conserv Biol J Soc Conserv Biol. avr 2019;33(2):434-43. doi:10.1111/cobi.13231 PubMed PMID: 30285277.

82. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. Res Synth Methods. sept 2019;10(3):330-42. doi:10.1002/jrsm.1354 PubMed PMID: null; PubMed Central PMCID: PMC6771536.

83. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. 26 janv 2010;11:55. doi:10.1186/1471-2105-11-55 PubMed PMID: 20102628; PubMed Central PMCID: PMC2824679.

84. Sercombe J, Bryant Z, Wilson J. Evaluating a Customized Version of ChatGPT for Systematic Review Data Extraction in Health Research: Development and Usability Study. JMIR Form Res. 11 août 2025;9:e68666. doi:10.2196/68666 PubMed PMID: 40789147; PubMed Central PMCID: PMC12338963.

85. Nussbaumer-Streit B, Ellen M, Klerings I, Sfetcu R, Riva N, Mahmić-Kaknjo M, et al. Resource use during systematic review production varies widely: a scoping review. J Clin Epidemiol. 1 nov 2021;139:287-96. doi:10.1016/j.jclinepi.2021.05.019

86. Gue CCY, Rahim NDA, Rojas-Carabali W, Agrawal R, Rk P, Abisheganaden J, et al. Evaluating the OpenAI's GPT-3.5 Turbo's performance in extracting information from scientific articles on diabetic retinopathy. Syst Rev. 16 mai 2024;13(1):135.

doi:10.1186/s13643-024-02523-2 PubMed PMID: 38755704; PubMed Central PMCID: PMC11097400.

87. Yao Y, Shen J, Luo J, Li N, Liao X, Zhang Y. Disagreements in risk of bias assessment for randomized controlled trials in hypertension-related Cochrane reviews. Trials. 21 juin 2024;25(1):405. doi:10.1186/s13063-024-08145-2 PubMed PMID: 38907276; PubMed Central PMCID: PMC11191165.

88. Jardim PSJ, Rose CJ, Ames HM, Echavez JFM, Van de Velde S, Muller AE. Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. BMC Med Res Methodol. 8 juin 2022;22:167. doi:10.1186/s12874-022-01649-y PubMed PMID: 35676632; PubMed Central PMCID: PMC9174024.

89. Armijo-Olivo S, Ospina M, da Costa BR, Egger M, Saltaji H, Fuentes J, et al. Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane risk of bias tool in physical therapy trials. PloS One. 2014;9(5):e96920. doi:10.1371/journal.pone.0096920 PubMed PMID: 24824199; PubMed Central PMCID: PMC4019638.

90. Minozzi S, Dwan K, Borrelli F, Filippini G. Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. J Clin Epidemiol. janv 2022;141:99-105. doi:10.1016/j.jclinepi.2021.09.021 PubMed PMID: 34537386.

91. Taneri PE. Human Versus Artificial Intelligence: Comparing Cochrane Authors' and ChatGPT's Risk of Bias Assessments. Cochrane Evid Synth Methods. sept 2025;3(5):e70044. doi:10.1002/cesm.70044 PubMed PMID: 40908961; PubMed Central PMCID: PMC12407253.

92. Crocker TF, Lam N, Jordão M, Brundle C, Prescott M, Forster A, et al. Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review. J Clin Epidemiol. sept 2023;161:39-45. doi:10.1016/j.jclinepi.2023.06.015 PubMed PMID: 37364620.

93. Zhang Y, Huang L, Wang D, Ren P, Hong Q, Kang D. The ROBINS-I and the NOS had similar reliability but differed in applicability: A random sampling observational studies of systematic reviews/meta-analysis. J Evid-Based Med. mai 2021;14(2):112-22. doi:10.1111/jebm.12427 PubMed PMID: 34002466.

94. Alaniz L, Vu C, Pfaff MJ. The Utility of Artificial Intelligence for Systematic Reviews and Boolean Query Formulation and Translation. Plast Reconstr Surg Glob Open. 30 oct 2023;11(10):e5339. doi:10.1097/GOX.0000000000005339 PubMed PMID: 37908326; PubMed Central PMCID: PMC10615538.

95. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? [Internet]. arXiv; 2023 [cité 10 oct 2025]. Disponible sur: http://arxiv.org/abs/2302.03495 doi:10.48550/arXiv.2302.03495

96. Cai X, Geng Y, Du Y, Westerman B, Wang D, Ma C, et al. Utilizing ChatGPT to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation [Internet]. medRxiv; 2025 [cité 10 oct 2025]. p. 2023.09.06.23295072. Disponible sur: https://www.medrxiv.org/content/10.1101/2023.09.06.23295072v2 doi:10.1101/2023.09.06.23295072

97. Trad F, Yammine R, Charafeddine J, Chakhtoura M, Rahme M, El-Hajj Fuleihan G, et al. Streamlining systematic reviews with large language models using prompt engineering and retrieval augmented generation. BMC Med Res Methodol. 10 mai 2025;25(1):130. doi:10.1186/s12874-025-02583-5

98. Vallamchetla SK, Abdelkader O, Elnaggar A, Ramadan D, Islam Shourav MM, Riaz IB, et al. Do it faster with PICOS: Generative AI-Assisted systematic review screening. J Biomed Inform. 1 août 2025;168:104860. doi:10.1016/j.jbi.2025.104860

99. Huang J, Lai H, Zhao W, Xia D, Bai C, Sun M, et al. Large Language Model–Assisted Risk-of-Bias Assessment in Randomized Controlled Trials Using the Revised Risk-of-Bias Tool: Evaluation Study. J Med Internet Res. 24 juin 2025;27:e70450. doi:10.2196/70450 PubMed PMID: 40554779; PubMed Central PMCID: PMC12238788.

100. Kim HY. Statistical notes for clinical researchers: Type I and type II errors in statistical decision. Restor Dent Endod. août 2015;40(3):249-52. doi:10.5395/rde.2015.40.3.249 PubMed PMID: 26295030; PubMed Central PMCID: PMC4534731.

101. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev. 21 août 2024;13(1):219. doi:10.1186/s13643-024-02609-x

102. Dai ZY, Shen C, Ji YL, Li ZY, Wang Y. Accuracy of Large Language Models for Literature Screening in Systematic Reviews and Meta-Analyses [SSRN Scholarly Paper] [Internet]. Rochester, NY: Social Science Research Network; 2024 [cité 10 oct 2025]. Disponible sur: https://papers.ssrn.com/abstract=4943759 doi:10.2139/ssrn.4943759

103. 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. 1 oct 2025;30(5):313-22. doi:10.1136/bmjebm-2024-113066 PubMed PMID: 40199559.

104. Li L, Sleem L, Gentile N, Nichil G, State R. Exploring the Impact of Temperature on Large Language Models:Hot or Cold? [Internet]. arXiv; 2025 [cité 27 févr 2026]. Disponible sur: http://arxiv.org/abs/2506.07295 doi:10.48550/arXiv.2506.07295

105. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. J Biomed Inform. 1 mars 2024;151:104620. doi:10.1016/j.jbi.2024.104620

106. Fagerberg P, Sallander O, Patil KV, Berg A, Nyman A, Borg N, et al. Dual-Model LLM Ensemble via Web Chat Interfaces Reaches Near-Perfect Sensitivity for Systematic-Review Screening: A Multi-Domain Validation with Equivalence to API Access [Internet]. medRxiv; 2025 [cité 14 août 2026]. p. 2025.11.03.25339455. Disponible sur: https://www.medrxiv.org/content/10.1101/2025.11.03.25339455v1 doi:10.1101/2025.11.03.25339455

107. Bramer WM, de Jonge GB, Rethlefsen ML, Mast F, Kleijnen J. A systematic approach to searching: an efficient and complete method to develop literature searches. J Med Libr Assoc JMLA. oct 2018;106(4):531-41. doi:10.5195/jmla.2018.283 PubMed PMID: 30271302; PubMed Central PMCID: PMC6148622.

108. Claude Opus 4.8 [Large language model] [Internet]. Anthropic; 2026. Disponible sur: https://claude.ai

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials1.1FullMethodology.docx
- SupplementaryMaterials1.2Fullprompts.docx
- SupplementaryMaterials2.1Searchstringresults.docx
- SupplementaryMaterials2.2Articlescreeningresults.docx
- SupplementaryMaterials2.3Dataextractionresults.docx
- SupplementaryMaterials2.4ROBassessmentresults.docx
- SupplementaryMaterials2.5.1Datasynthesisresults.docx
- SupplementaryMaterials2.5.2Authors.xlsx
- SupplementaryMaterials2.5.2GPT.xlsx
- SupplementaryMaterials3.1Stepspecificchallenges.docx