BMJ Digital Health & AI

Large language model enhanced framework for systematic reviews and meta-analyses

Jiashu Shen,[1] Zhiyao Luo,[2] Danping Jia,[3] Siran Wang,[4] Feng Sun,[5] Jinge Wu [6]

To cite: Shen J, Luo Z, Jia D, *et al*. Large language model enhanced framework for systematic reviews and meta-analyses. *BMJ Digital Health and AI* 2025;**1**:e000017. doi:10.1136/bmjdhai-2025-000017

Received 3 February 2025
Accepted 6 August 2025

[1]Nuffield Department of Medicine, University of Oxford, Oxford, UK
[2]Department of Engineering Science, University of Oxford, Oxford, UK
[3]Bayer (China) Limited, Beijing, China
[4]Analytics, Technology and Operations Department, University of Leeds, Leeds, UK
[5]Peking University, Beijing, China
[6]Institute of Health Informatics, University College London, London, UK

Correspondence to
Dr Jinge Wu;
jinge.wu.20@ucl.ac.uk

ABSTRACT

Objective To evaluate and synthesise current applications of large language models (LLMs) in systematic reviews and meta-analyses (SRMAs), identify key limitations and propose an enhanced theoretical framework to improve the efficiency, scalability and reliability of evidence synthesis.

Methods and analysis We conducted a narrative review of recent studies applying LLMs across key SRMA stages. A total of 21 publications were analysed for model type, task application, accuracy metrics and workflow impact. Building on this evidence base, we designed a comprehensive LLM-enhanced SRMA framework that categorises LLM roles as consultants and assistants, integrates human-in-the-loop strategies and uses retrieval-augmented generation (RAG) and agent-based architectures to address critical challenges including hallucinations, bias and workflow inefficiency.

Results The reviewed literature demonstrated that LLMs can support various SRMA tasks with reported accuracy ranging from 61% to 99%, showing particular promise in literature screening and data extraction. Our proposed framework conceptualises modular integration of LLMs across all six SRMA stages, with LLMs serving as consultants for research question formulation and search strategy development and as assistants for task automation including abstract screening and structured data extraction. The framework incorporates RAG technology to reduce hallucinations by grounding outputs in retrieved literature and employs agent-based orchestration for complex analytical workflows. Theoretical analysis suggests potential for significant efficiency gains while maintaining methodological rigour through strategic human oversight.

Conclusion LLMs offer substantial theoretical potential to transform evidence synthesis by improving efficiency, scalability and consistency across SRMA workflows. The proposed LLM-enhanced framework provides a systematic, theoretically grounded approach for integrating advanced artificial intelligence capabilities into existing SRMA methodologies while preserving essential human oversight and analytical integrity. Future empirical studies are needed to validate the framework's practical effectiveness, establish implementation protocols and demonstrate real-world benefits in evidence-based medicine.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Systematic reviews and meta-analyses (SRMAs) are time-consuming and demand substantial human effort.
⇒ Large language models (LLMs) have demonstrated potential in automating discrete SRMA tasks, such as literature screening and data extraction.

WHAT THIS STUDY ADDS

⇒ This study introduced a comprehensive, modular framework that integrated LLMs throughout all stages of the SRMA process, assigning them roles as consultants and assistants.
⇒ It outlined detailed implementation strategies with concrete examples and evaluated the performance, reliability and practical considerations of LLMs within this structured approach.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The proposed framework has the potential to substantially reduce the labour and time involved in producing high-quality SRMAs, while maintaining transparency and methodological rigour.
⇒ It offers a practical blueprint for integrating LLMs into SRMA workflows and supports the development of living, continuously updated reviews.

INTRODUCTION

Systematic reviews and meta-analyses (SRMAs) are regarded as the most rigorous methods for gathering and synthesising findings from diverse clinical studies to evaluate the effectiveness of interventions and inform clinical guidelines.[1 2] This comprehensive approach increases statistical power and diminishes bias, elevating the findings' accuracy and reliability[3] and positioning them at the top of the clinical evidence hierarchy.[1]

The process of SRMAs mainly involves six principal steps: (1) research question formulation, (2) literature searching, (3) study selection, (4) data extraction, (5) quality assessment and (6) data analysis and interpretation.[4–6] However, this approach encounters several challenges. Primarily,

the performance of SRMAs is based on the expertise of human researchers, which is inherently limited in scope and subject to human error and bias. Furthermore, while semiautomated tools such as Covidence, Rayyan, DistillerSR and ASReview, as well as AI-powered discovery platforms like Scite.ai, Undermind and ResearchRabbit, have been developed to support literature screening, exploration and data management, traditional SRMAs remain largely labour-intensive. They still require at least two independent reviewers to manually screen thousands of papers and extract clinical data. Recent studies have systematically evaluated the application of large language models (LLMs) in healthcare evidence synthesis, highlighting their potential to automate literature screening, data extraction and bias assessment in SRMAs.[7–9] A typical Cochrane SRMA demands a research team to invest 2–3 years from protocol inception to final results publication.[10] This lengthy process also leads to the near-immediate obsolescence of published SRMA findings, particularly in rapidly evolving domains. In addition, conducting SRMAs demands substantial resources and funding, with estimates from 2019 suggesting that each SRMA costs approximately US$141 195.[11] For each research university in the USA, the cumulative annual expenditure for SRMA reaches an average of up to US$18 660 305, while for each pharmaceutical company, it amounts to US$16 761 235.[11] The growing number of SRMA publications underscores the urgent necessity for an efficient and innovative approach.[12]

The emergence of LLMs, a type of generative artificial intelligence (generative AI)[13] specialising in text-based data, presents transformative solutions for SRMAs.[14 15] Unlike conventional machine learning techniques, LLMs like OpenAI's GPT series model (which powers such as GPT, InstructGPT[16] and ChatGPT) and Google's Gemini[17] leverage deep learning and massive data sets to enable advanced language understanding and generation.[18] LLMs excel in various natural language processing (NLP) tasks and can demonstrate substantial knowledge and comprehension in medical domains.[19 20] There are some preliminary attempts to use LLMs to address individual steps in SRMAs.[21–25] However, currently, there is no standard protocol for integrating LLMs into the existing SRMA system. This study aims to review the previous studies in applying LLMs to SRMA processes and introduce a systematic LLMs-enhanced framework empowered by LLMs, retrieval-augmented generation (RAG) and agents, intended to stimulate discussion and guide future methodological development. In addition, we propose a collaborative model between LLMs and researchers for conducting SRMAs, fostering synergy and optimising outcomes.

REVIEW OF LARGE LANGUAGE MODELS IN SYSTEMATIC REVIEWS AND META-ANALYSES

Table 1 provides an overview of the current applications and performance of LLMs in the specific steps of SRMAs. It summarised 15 publications, detailing each study's application steps and sample size, as well as the specific LLM models and approaches used (eg, Prompt, RAG, Agent). We evaluated the accuracy of LLMs in different research contexts, ranging from 61% to 99%. The 'grading for efficacy improvement' was performed independently by two reviewers, based on qualitative evaluation of reported outcomes in terms of screening or extraction accuracy, efficiency gains and overall contribution to workflow enhancement. Overall, the table clearly illustrated the trends and potential benefits of employing LLMs for large-scale data screening and textual analysis in SRMAs. Most studies reported a significant improvement in screening efficiency and accuracy when LLMs were incorporated.

Capabilities of LLMs across key SRMA workflow stages. We review the literature and generate a table which compares the strengths and limitations of traditional SRMAs versus LLM-enhanced SRMA across various aspects, including accuracy, analytical capabilities, bias, comprehensiveness, consistency, efficiency and ethical transparency concerns (table 2). Traditional SRMA generally offers higher analytical capabilities and transparency but is prone to human error and subjective biases and is time-consuming due to manual processes. In contrast, LLM-enhanced SRMA may be able to increase efficiency and provide broader contextual searches; however, it faces challenges such as limited statistical analysis capabilities, inherited biases from training data and opaque decision-making processes that may reduce transparency and consistency. The comparison highlights that while LLMs can complement traditional methods by automating certain stages, they also introduce novel risks that must be carefully managed.

We summarised the functional roles of LLMs across six key stages of the SRMA process in table 3. LLMs serve as either conceptual advisors or operational assistants, depending on the task. For example, they support Patient, Intervention, Comparison and Outcome (PICO) formulation by synthesising prior findings, enhancing literature searching through semantic query rewriting and having the potential to assist in classifying study eligibility based on inclusion criteria. LLMs also aid in extracting structured data (eg, sample sizes, effect estimates), interpreting quality appraisal tools and generating narrative summaries of findings. However, statistical analyses are performed using conventional tools. Across all stages, human oversight remains critical to ensure accuracy and transparency.

TECHNICAL FOUNDATION

Large language models

LLMs, such as GPT-4, are transformer-based deep learning models trained on large-scale corpora to generate and interpret natural language. LLMs adopt deep learning techniques and are trained using architectures with billions to hundreds of billions of parameters, leveraging

Table 1 Applications and performance of large language models in systematic reviews and meta-analyses

Publications	Research question formulation	Study selection			Quality assessment	Data extraction	Data analysis and interpretation	
		Literature searching	Titles/abstracts screening	Full texts screening			Data analysis	Interpretation
Adam *et al*[67]		10346						
Syriani *et al*[45]		5144						
Alshamii *et al*[68]		496	496	145				
Li *et al*[69]			505					
Gargari *et al*[22]			330					
Guo *et al*[70]			24000					
Li *et al*[69]			2					
Issaiy *et al*[71]			1198					
Cai *et al*[72]			8550					
Delgado-Chaves *et al*[52]			6469					
Luo *et al*[73]			24534	4096				
Khraisha *et al*[74]			300	150		30		
Chen *et al*[75]			1086					
Omai *et al*[76]			20 170					
Wang *et al*[77]				50 000				
Hasan *et al*[78]					307			
Mahuli *et al*[79]					1	1		
Kartchner *et al*[80]						2 data sets		
Shah-Mohammadi and Finkelstein[81]						1		
Reason *et al*[24]						4	4	4
Lam-Hoai and Simonart[82]								16

Publications	LLM model	LLM approach	Accuracy	Grading for efficacy improvement
Adam *et al*[67]	GPT-4	Prompt	85%	High
Syriani *et al*[45]	GPT 3.5	Prompt	82%	High
Alshamii *et al*[68]	GPT-3.5	Prompt	88%	High
Li *et al*[69]	9 models	Prompt	71–96%	High
Gargari *et al*[22]	GPT-3.5 Turbo	Prompt, RAG	82% in the best prompt	High

Continued

Table 1 Continued

Publications	LLM model	LLM approach	Accuracy	Grading for efficacy improvement
Guo *et al*[70]	GPT-4	Prompt	91%	High
Li *et al*[69]	GPT-4	Prompt	84–95%	High
Issaiy *et al*[71]	GPT-3.5 Turbo	Prompt	95%	High
Cai *et al*[72]	GPT-3.5 and GPT-4	Prompt	81–87%	High
Delgado-Chaves *et al*[52]	18 LLM models	Prompt	40–92%	High
Luo *et al*[73]	GPT-3.5 Turbo	Prompt	80–95%	High
Khraisha *et al*[74]	GPT-4	Prompt	Screening: 54–96%; Extraction: 81–85%	High
Chen *et al*[75]	GPT-4	Prompt	33%	Medium
Omai *et al*[76]	GPT-4 Turbo	Prompt	75%	High
Wang *et al*[77]	5 models	Prompt, calibration, ensemble	Depends on data sets and models: 48–75%	High
Hasan *et al*[78]	GPT-4	Prompt, RAG, agent	61%	Medium
Mahuli *et al*[79]	GPT-3.5	Prompt	Not provided	High
Kartchner *et al*[80]	GPT 3.5 Turbo and GPT-JT	Prompt	Depends on the extracted term: 20–90%	High
Shah-Mohammadi and Finkelstein[81]	GPT-3.5	Prompt	Depends on the prompts: 44–97%	High
Reason *et al*[24]	GPT-4	Prompt	99%	High
Lam-Hoai and Simonart[82]	GPT-3.5	Prompt	75%	Medium

The green-shaded cells indicate the specific steps in the systematic review and meta-analysis process where the corresponding studies applied large language models. The numbers within the green cells represent the sample sizes associated with each step.
LLM, large language model; RAG, retrieval-augmented generation.

Table 2 Comparison of traditional and large language model-enhanced systematic reviews and meta-analyses

Aspect with references	Traditional systematic reviews and meta-analyses	LLM-enhanced systematic reviews and meta-analyses
Accuracy[83 84]	⚠ Prone to human error and inconsistency: data extraction and interpretation are susceptible to subjective errors and bias.	⚠ Risk of hallucinations and misinterpretations: LLMs may generate inaccurate information or fabricate study details, affecting review reliability.
Analytical capabilities[7 85]	✓ High analytical capabilities with advanced statistical methods: requires specialised tools and statistical knowledge for complex analysis.	⚠ Limited statistical analysis capabilities: useful for preliminary analyses and summarisation but lacks the depth of specialised statistical tools.
Bias[86]	⚠ Prone to subjective bias and selection bias: human reviewers may unintentionally favour certain outcomes or overlook key studies.	⚠ Inherits biases from training data: LLMs reflect biases present in training data sets and may lack appropriate context for specific domains.
Comprehensiveness[87]	⚠ Limited by keyword search and predefined criteria: restricted by search keywords, language and scope.	✓ Border and context-aware search: leverages broader contextual understanding, supports multilingual literature and captures nuances beyond keywords.
Consistency[88 89]	⚠ Variable and subjective consistency: consistency can vary across different reviewers and their interpretation of studies.	⚠ Inconsistent performance across domains: model performance may vary depending on the nature of training data, leading to domain-specific inconsistencies.
Efficiency[90]	✗ **Time-consuming and labour-intensive**: manual processes for screening, data extraction and synthesis are slow and resource-heavy.	✓ High efficiency and automation: automates various stages like literature screening, data extraction and initial synthesis.
Ethical and transparency concerns[66]	✓ Transparent and traceable decision-making: human decision processes are easier to trace and explain.	⚠ Opaque decision-making and potential ethical issues: model decisions lack transparency, and it may be difficult to trace the rationale behind certain inclusions or exclusions.

✓ = LLM typically supports this stage; ⚠ = Requires careful human review; ✗ = Not suitable for LLM automation.
LLM, large language model.

data sets comprising trillions of tokens. This vast scale enables them to understand, summarise, generate and predict complex language content.[26] Unlike traditional NLP models, LLMs are developed with advanced architectures, such as transformers, which are capable of understanding contextual information.[27] Typical LLM training comprises three stages: pretraining, supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF).[28] During pretraining, models acquire an understanding of language patterns from large-scale textual data, while SFT adapts them to specific tasks using smaller, targeted data sets.[29] RLHF further

Table 3 Capabilities of large language models (LLMs) across systematic reviews and meta-analyses workflow stages.

SRMA stage	Primary LLM role	Role type	Function
Research question formulation	Conceptual advisor	Consultant	Identifies knowledge gaps, formulates PICO elements, refines scope
Literature searching	Semantic rewriter and optimiser	Consultant	Suggests synonyms, expands queries, compares search strategies based on retrieved outputs
Study selection	Triage assistant	Assistant	Classifies abstracts (eg, include/exclude/unsure) with rationale; prioritises borderline studies
Data extraction	Structured extractor	Assistant	Extracts key variables (eg, sample size, effect sizes, CIs) from text/tables
Quality assessment	Criteria interpreter	Assistant	Support in applying appraisal tools, identification of limitations
Data analysis and interpreting	Summariser and explainer	Assistant	Narrative summaries of findings, highlighting patterns; delegates statistics to scripts

PICO, Patient, Intervention, Comparison and Outcome; SRMA, systematic reviews and meta-analysis.

enhances generative capabilities through human feedback.[30] In application, LLMs exhibit zero-shot learning, allowing them to perform tasks without explicit training and few-shot learning, where they generalise from a few provided examples. These techniques enable LLMs to be applied across various industries for tasks such as content generation, data mining and knowledge retrieval.[18 31]

In the proposed framework, LLMs are positioned as consultants and assistants to enhance different stages of the SRMA process. In the consultant role, LLMs support early-stage tasks such as formulating research questions, refining eligibility criteria and structuring search strategies. By leveraging their extensive knowledge base, LLMs assist researchers in conceptualising study designs and ensuring comprehensive coverage of the research domain.

In the assistant role, LLMs contribute to task execution throughout the SRMA workflow. This role operates in two modes: automating non-critical steps, such as literature retrieval, deduplication and preliminary data extraction; and supporting critical steps, including quality assessment and risk of bias evaluation. For critical tasks, LLMs assist by extracting relevant information and structuring preliminary evaluations, while final decisions are made by human reviewers to ensure methodological rigour. Through this dual-role framework, LLMs enhance workflow efficiency, uphold research integrity and promote scalable, transparent evidence synthesis.

Retrieval-augmented generation

RAG[32] stands as a promising technique to enhance the efficacy of LLMs within the domain of SRMAs.[33] RAG is capable of retrieving pertinent information from external knowledge repositories, such as scientific literature.[34] The process involves preprocessing medical literature by converting textual information into vector embeddings—numerical representations that capture semantic meaning. These embeddings map documents into a high-dimensional space where semantically similar studies are positioned closer together. When a new task arises, the system retrieves the most relevant documents based on proximity in the embedding space, providing accurate and contextually grounded information to augment the LLM's responses.[35–37] We incorporate RAG to enhance factual accuracy and domain-specific reasoning by dynamically retrieving external documents (eg, abstracts, clinical trial reports or guidelines) and feeding them into the LLM's prompt context. This not only grounds the model's outputs in verifiable, up-to-date source material—especially valuable for tasks such as full-text screening, evidence synthesis and structured summarisation—but also enables the model to reason over specialised or under-represented topics (eg, rare diseases, newly approved drugs) without requiring additional training. In doing so, RAG mitigates hallucinations and reduces the need for costly and frequent model fine-tuning.

In the SRMA processes, researchers face the challenge of accurately extracting and synthesising information from numerous studies. While LLMs are powerful, they may sometimes produce inaccurate responses or 'hallucinations'. RAG mitigates this risk by grounding LLM outputs in retrieved evidence, improving the accuracy and reliability of generated insights.[38]

Agent

Agent represents a sophisticated approach within generative AI, combining LLMs with a suite of tools to accomplish intricate tasks.[39] In the context of SRMAs, an agent can be designed as a software system that harnesses LLMs and various functional tools to automate and streamline the review process.[40] For tasks such as data analysis and interpretation, an agent can provide support from the following aspects:

1. Task decomposition: The agent breaks down the given task into smaller, manageable subtasks, such as data processing, analysis and interpretation.
2. Tool selection and integration: The agent programmatically defines and selects tools for each subtask. For instance, data processing functions can use data packages, analysis functions can leverage statistical libraries and interpretation functions can use RAG.
3. Input and output design: The agent outlines input requirements and expected outputs for each subtask, ensuring that data flows smoothly between different stages of the SRMA process.
4. Task orchestration: The agent coordinates subtasks, manages data flow and integrates results from various stages to complete the overall task efficiently.

Researchers can optimise various facets of the SRMA workflow by leveraging an LLM agent, reducing time spent on labour-intensive tasks like literature search and data extraction. This allows researchers to focus on higher-level analysis and interpretation, ultimately enhancing SRMA efficiency and quality. Combining LLMs' strengths with specialised tools, the agent furnishes researchers with a comprehensive and intelligent platform for conducting SRMA.

A LARGE LANGUAGE MODEL-ENHANCED SYSTEMATIC REVIEWS AND META-ANALYSES FRAMEWORK

The Cochrane Handbook[41] serves as a widely recognised standard for conducting SRMAs, providing comprehensive methodologies to ensure rigour, transparency and minimisation of bias. Drawing on its guidelines, we developed an LLM-enhanced framework for SRMAs, as illustrated in figure 1. The SRMA process is organised into six principal steps: (1) research question formulation, (2) literature searching, (3) study selection, (4) data extraction, (5) quality assessment and (6) data analysis and interpretation. For each step, we reviewed traditional manual approaches, identified their limitations and proposed how these tasks could be supported and optimised through the integration of LLMs.

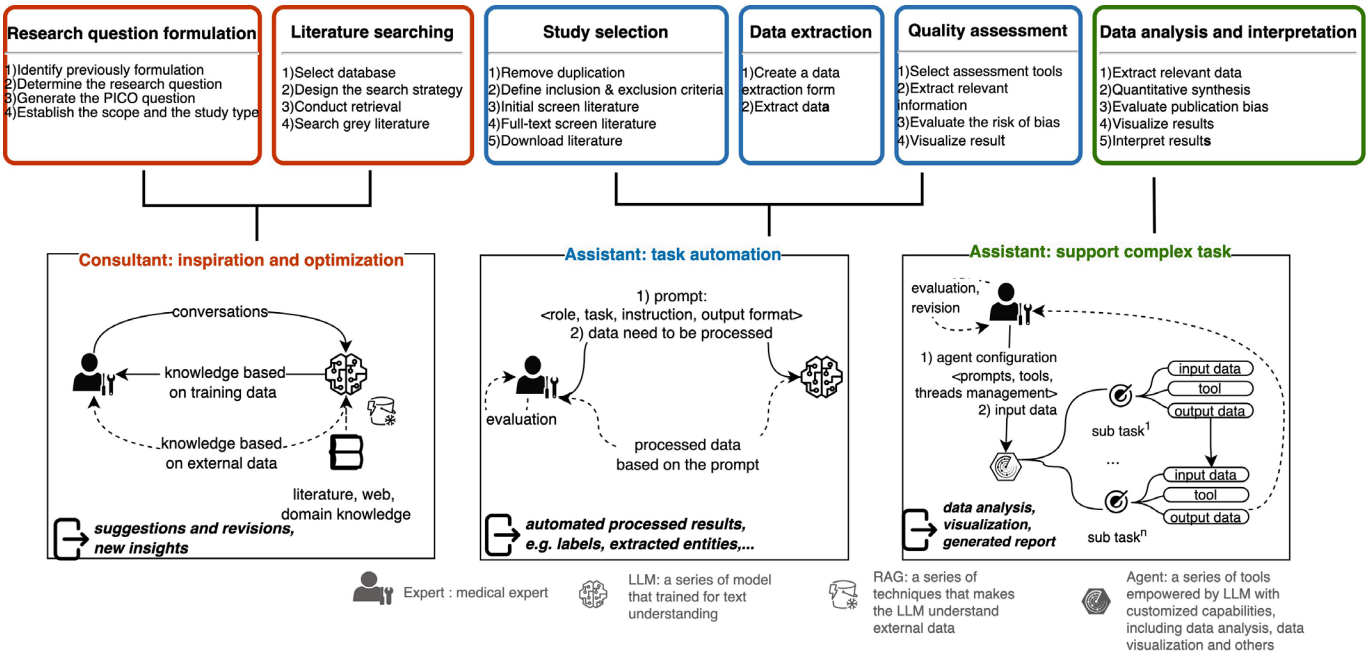

Figure 1 A large language model-enhanced systematic reviews and meta-analyses framework. The figure illustrates the consultant and assistant roles of LLMs and the integration of RAG and agent modules across the SRMA workflow. In the framework of LLM-enhanced SRMAs, we formulate the role of LLM and related techniques as the role of consultant and assistant. The consultant leverages LLM and RAG to provide inspiration and optimisation for cognitively demanding tasks, such as research question formulation and literature searching. The assistant is divided into two types: task automation and complex task support. The task automation assistant focuses on supporting labour-intensive steps, including study selection, data extraction and quality assessment. In contrast, the complex task support assistant leverages agent technologies to assist with data analysis-related tasks with the power of agent techniques. LLM, large language model; PICO, Patient, Intervention, Comparison and Outcome; RAG, retrieval-augmented generation; SRMA, systematic reviews and meta-analysis.

Research question formulation

Formulating a research question is a critical step in conducting SRMAs, as it ensures that research efforts address answerable questions and fill important knowledge gaps.[41] Traditionally, researchers develop the research question and PICO framework based on literature reviews and the identification of unmet needs. However, this process may be influenced by information silos and cognitive biases.[42]

LLMs may be able to assist by searching published SRMAs, systematic reviews and clinical guidelines to map the existing research landscape. By integrating with techniques such as semantic clustering and topic modelling—where *semantic clustering* refers to grouping documents based on vector similarity in embedding space, and *topic modelling* (eg, Latent Dirichlet Allocation) uncovers latent thematic structures—LLMs have the potential to assist in identifying frequently investigated topics, emerging trends and areas with limited evidence. While topic modelling itself is not a native function of LLMs, they can contribute by generating or structuring inputs for such analyses. Based on this analysis, LLMs may be able to propose refined research questions aligned with PICO structures and the objectives of the planned SRMA. LLMs also help optimise the clarity and feasibility of clinical questions. Final selection and refinement are performed by human researchers to ensure methodological rigour and relevance.

Illustrative example: LLM application in research question formulation

Our framework applies LLMs to research question formulation through a structured process using GPT-4 via the OpenAI API, combining interactive prompting and programmatic queries with few-shot examples. Abstracts were retrieved from PubMed using E-utilities, and the system applied sentence embedding-based similarity (using sentence-transformers, eg, all-MiniLM-L6-v2) followed by k-means clustering to identify thematic research gaps (eg, limited evidence for primary prevention in patients without pre-existing cardiovascular (CV) disease (CVD)). Human researchers reviewed and validated all model-suggested PICO elements. While GPT-4 was used in this implementation, our framework is model-agnostic and can be readily adapted to other commercial APIs (eg, Claude, Gemini, DeepSeek) or open-source LLMs (eg, LLaMA, Mistral, Qwen), depending on access, performance and domain-specific requirements.

Using antidiabetic medications as an example, the LLM analyses literature on 'SGLT2 inhibitors' and 'GLP-1 receptor agonists' for CV outcomes from PubMed. The system processes retrieved abstracts using GPT-4, then applies semantic grouping based on population, intervention and outcomes. LLMs summarise limitations and future directions mentioned in conclusions of prior studies, such as lack of trials in low-risk populations. These insights inform identification of evidence

gaps (eg, limited data for primary prevention in patients without established CVD), from which the LLM proposes targeted PICO elements. For example: 'limited evidence for primary prevention in patients without pre-existing cardiovascular disease'. Based on this analysis, the LLM generates targeted PICO frameworks: P (patients with T2DM without established CVD), I (SGLT2 inhibitors), C (GLP-1 receptor agonists), O (heart failure hospitalisation). This method could potentially reduce question formulation time from weeks to days through systematic knowledge gap analysis. Future studies will need to evaluate the comprehensiveness and clinical relevance of LLM-identified research questions compared with expert-formulated ones.

Literature searching

Literature searching in SRMAs requires developing a structured strategy based on predefined keywords and querying appropriate databases such as PubMed and Embase.[41] Grey literature sources, including OpenGrey and ProQuest Dissertations and Theses, are also consulted to reduce bias, enhance transparency and ensure comprehensive coverage.[43] A robust and systematic search forms the foundation for the validity of subsequent analyses.

LLMs may be able to assist by identifying key elements from research questions (eg, PICO components), expanding search terms with synonyms, abbreviations and controlled vocabularies such as Medical Subject Headings (MeSH) and suggesting Boolean search structures to optimise sensitivity and specificity. Training LLMs across diverse languages and cultural contexts further enhances the global relevance of evidence synthesis.[44]

LLMs may assist by identifying key elements from research questions—typically framed using structures such as PICO—and expanding search terms with relevant synonyms, abbreviations and controlled vocabularies (eg, MeSH, Emtree). They also aid in constructing Boolean queries that combine concepts logically using operators (AND, OR, NOT), while adjusting syntax based on specific database requirements. Additionally, LLMs can iteratively evaluate search results and propose refinements to improve precision and recall. Training LLMs across diverse languages and cultural contexts further enhances the global relevance of evidence synthesis.

The feasibility of database access must also be considered. While PubMed offers open API access supporting automated interactions, other major databases such as MEDLINE (via Ovid), EMBASE and Web of Science require institutional subscriptions and enforce licensing restrictions that limit direct querying by LLMs. In most current applications, comprehensive searches still require human researchers to manually access licensed platforms. However, it is technically feasible to embed LLMs within systems that authenticate via institutional credentials, enabling programme access to subscription databases. With proper licensing and security protocols, this setup could allow LLMs to operate within the researcher's access privileges, significantly expanding their utility in automated evidence retrieval. Manual review and transparent documentation of the search process remain essential to ensure accuracy, reproducibility and legal compliance.

Illustrative example: LLM application in literature searching

Building on the SGLT2 inhibitor versus GLP-1 receptor agonist research question for CV outcomes, we used GPT-4 (OpenAI API) to assist with term expansion, MeSH mapping and Boolean query construction. The LLM-generated queries were tested against PubMed's E-utilities API. Although the model generated initial strategies, all final queries were refined and validated by researchers to ensure database compatibility and comprehensiveness. The system first extracts PICO elements from our research question: P (patients with T2DM without established CVD), I (SGLT2 inhibitors), C (GLP-1 receptor agonists), O (heart failure hospitalisation). The LLM expands these terms using biomedical knowledge. For example, 'SGLT2 inhibitors' is expanded to include MeSH terms, specific drug names (empagliflozin, dapagliflozin, canagliflozin) and brand names (Jardiance, Farxiga, Invokana). Similarly, GLP-1 terms include class terms and specific agents (semaglutide, liraglutide, dulaglutide). The LLM then constructs a precise Boolean strategy such as: ("Diabetes Mellitus, Type 2"[Mesh] OR "type 2 diabetes"[tiab] OR T2DM[tiab]) AND ("Sodium-Glucose Transporter 2 Inhibitors"[Mesh] OR "SGLT2 inhibitor"[tiab] OR empagliflozin[tiab] OR dapagliflozin[tiab]) AND ("GLP-1 Receptor Agonists"[Mesh] OR "GLP-1 agonist"[tiab] OR semaglutide[tiab] OR liraglutide[tiab]) AND ("Heart Failure"[Mesh] OR "heart failure hospitalization"[tiab] OR "cardiovascular outcomes"[tiab]). When initial search results were sparse or missed key studies, the system flagged them as insufficient. The LLM (GPT-4) then proposed alternative terms and reformulated queries using semantic reasoning. Multiple strategies were generated with varied prompts and tested via public APIs (eg, PubMed) using standard scripts that assessed result counts and overlaps. Human reviewers evaluated the outputs for relevance, coverage and inclusion of sentinel studies. The most effective strategy was selected through this iterative, hybrid process, combining LLM-driven language reasoning with the reliability of deterministic code.

Study selection

Study selection aims to identify the most relevant literature based on predefined inclusion and exclusion criteria. Traditionally, experts screen the literature by examining metadata, abstracts and, when necessary, full texts from multiple perspectives.[41] This process is labour-intensive and time-consuming. Recent studies have shown that LLMs have the potential to assist in study selection with accuracy comparable to traditional manual screening.[22 45] LLMs may also help balance the

coverage-precision trade-off through self-evaluation mechanisms, improving both comprehensiveness and specificity.[46 47]

In our framework, LLMs assist with prescreening by classifying abstracts as 'include', 'exclude' or 'uncertain' using structured prompts aligned with eligibility criteria. Unlike tools such as Covidence, a screening platform, which rely on feedback-driven ranking without rationale, LLMs provide justifications and can deprioritise clearly irrelevant studies while highlighting borderline cases. Final decisions remain with human reviewers, and discrepancies—whether with the LLM or between reviewers—are resolved through consensus or third-party adjudication. This collaborative workflow improves screening efficiency while preserving methodological rigour, transparency and reproducibility.[48]

Illustrative example: LLM application in study selection

For our antidiabetic medication SRMA, the study selection component uses a three-tier classification system. The system implements a prompt structure including: (1) explicit eligibility criteria based on the PICO framework; (2) exemplars of clearly eligible studies (randomised controlled trials (RCTs) comparing SGLT2i vs GLP-1RA with heart failure outcomes), clearly ineligible studies (observational studies, wrong comparators) and borderline cases with reasoning; and (3) classification instructions requiring extraction of supporting evidence from the text. The LLM processes abstracts and categorises them as 'definitely include', 'definitely exclude' or 'uncertain' requiring human review. 'Definitely exclude' primarily comprises wrong comparator or study design papers, while 'definitely include' contains clearly described eligible RCTs. This tiered approach reduces screening workload while maintaining high recall. Future validation studies will optimise confidence thresholds for different review types to maximise screening efficiency.

Data extraction

Data extraction is a critical step in SRMAs, requiring the systematic retrieval of information such as study design, participant characteristics, interventions and outcomes using standardised protocols. Traditionally performed by two independent reviewers, this process is time-consuming and resource-intensive.[49] LLMs may assist by applying techniques such as named entity recognition, relation extraction and table or figure parsing to identify relevant data from both textual and visual elements. Extracted information is synthesised into structured formats, such as spreadsheets or databases, facilitating efficient data management and statistical analysis.

To improve accuracy and reduce hallucinations, the framework employs few-shot prompting with domain-specific examples, structured output formats to constrain model generation and consistency checks across multiple LLM runs. Extracted terms are linked to standard biomedical vocabularies (eg, MeSH or Unified Medical Language System (UMLS)), enhancing entity grounding and reducing ambiguity. Where appropriate, traditional NLP tools such as MedCAT or cTAKES can be invoked as agents to support terminology disambiguation, entity linking or concept normalisation—particularly in high-stakes or ambiguous cases. To promote reproducibility, all prompts and extraction outputs are systematically archived and standardised templates are used across data sets. Disagreements between LLM outputs and human reviewers are resolved through adjudication by a second reviewer following prespecified resolution criteria. Multi-pass extraction with varied prompts and semantic similarity checks is also employed to reduce omissions. These safeguards collectively enhance the reliability, reproducibility and comprehensiveness of the LLM-assisted data extraction process in specialised medical domains.

Illustrative example: LLM application in data extraction

For the included antidiabetic medication studies, our framework implements a structured data extraction approach. We develop a hierarchical extraction schema specifying required data fields across six categories: study metadata, population characteristics (glycated haemoglobin, body mass index, CVD history), intervention details (SGLT2i drug, dosage, duration), comparison treatments (GLP-1RA specifics), outcomes (primary: heart failure hospitalisation; secondary: Major Adverse Cardiovascular Events (MACE), CV death, all-cause mortality) and risk of bias indicators. The LLM extraction pipeline used GPT-4 accessed via API and involved a three-stage approach: identification of relevant sections, structured extraction into predefined templates and self-verification via repeated prompt variants. While the model handled initial extraction, ambiguous cases—particularly numerical outcomes and adverse events—were flagged for human verification. For heart failure outcomes, the system enforces specific format requirements including intervention and control group events, totals, HRs and CIs. Testing shows the system accurately extracts most data points on first pass, with human verification focusing only on flagged uncertain fields, primarily complex statistical outcomes and adverse event reporting.

Quality assessment

During the quality assessment phase of a meta-analysis, researchers use standardised tools to evaluate methodological quality and risk of bias. For RCTs, the Cochrane Risk of Bias 2.0 (RoB 2.0) tool[50] assesses domains such as randomisation, deviations from intended interventions, missing outcome data and outcome measurement. For non-randomised studies, the Newcastle-Ottawa Scale (NOS)[51] evaluates selection, comparability and outcome ascertainment.

LLMs offer a novel approach for enhancing this process by automating the extraction and structuring of risk-relevant information. Rather than replacing expert judgement, LLMs function as decision-support tools that prescreen study texts, highlight reporting gaps and align extracted content with domains in RoB 2.0 or NOS. For

example, an LLM may identify descriptions of randomisation procedures, blinding or missing data in RCTs and evaluate cohort comparability or outcome assessment adequacy in observational studies. This capability enables a more efficient, consistent and scalable initial assessment, particularly when screening large volumes of studies. Recent research shows that LLM-assisted outputs can achieve accuracy comparable to human reviewers,[50–52] though final risk-of-bias judgments must remain under expert supervision to ensure methodological rigour.

Illustrative example: LLM application in quality assessment

For assessing the quality of the included antidiabetic medication trials, our framework implements quality assessment through domain-specific prompt engineering aligned with validated tools. For RoB 2.0 assessments, we create a structured prompt template with five explicit domains and specific extraction instructions. For example, in the 'randomization process' domain, the prompt instructs: 'Extract all text describing randomization method, sequence generation, and allocation concealment. If no information is provided, state 'Not reported". The system applies a sequential process: first extracting relevant text fragments, then generating an assessment ('Low risk of bias'), followed by a structured rationale citing specific RoB 2.0 criteria. We implement confidence scoring for each domain assessment, with low confidence triggering human review. Testing shows good agreement with expert assessment across domains, with lowest agreement in the 'deviations from intended interventions' domain where medication class crossovers complicate assessment. This approach maintains assessment quality while reducing time requirements, with human reviewers focusing on low-confidence domains and final judgement verification.

Data analysis and interpretation

Data analysis in SRMAs synthesises results through effect size calculation, heterogeneity analysis, subgroup analysis, sensitivity testing and bias assessment. Traditional software such as RevMan enables these steps but often requires manual operation and specialist knowledge. LLM agents offer an advanced approach by automating complex tasks. For example, an agent can automatically identify covariates from study characteristics and stratify studies for subgroup analyses, such as by age group or intervention dosage. Agents may also iteratively conduct sensitivity analyses by excluding high-bias studies and dynamically updating pooled effect estimates.[53] RAG further supports the interpretation of complex statistical outputs by grounding them in retrieved literature. While agent-based approaches improve the efficiency and robustness of data synthesis, human oversight remains essential to validate results and ensure the reliability of SRMA conclusions.[54]

Illustrative example: LLM application in data analysis

For analysing the comparative effects of SGLT2 inhibitors versus GLP-1 receptor agonists, our framework employs a modular agent architecture. The Analysis Planning Module uses a decision tree to identify appropriate analytical approaches based on data characteristics (heterogeneity triggers random-effects model; sufficient studies per antidiabetic class enable subgroup analysis; baseline CVD risk variation suggests meta-regression). The Statistical Processing Module executes effect size calculation, medication subgroup comparisons (empagliflozin vs semaglutide vs other agents) and meta-regression of treatment effect against baseline CVD risk. Results show significant heterogeneity for heart failure hospitalisation outcomes, with meta-regression identifying baseline CVD risk as explaining substantial variance. SGLT2 inhibitors show advantages for heart failure hospitalisation while GLP-1 receptor agonists demonstrate benefits for MACE outcomes. The Interpretation Module generates structured findings highlighting these differential benefits and detects publication bias for smaller trials. This agent-based approach significantly reduces analysis time while enabling comprehensive exploration of prespecified and data-driven subgroup analyses.

While this illustrative example primarily demonstrates technical feasibility and workflow integration, formal evaluation of performance is still underway. Preliminary internal testing suggests that the LLM-enhanced workflow reduced total analysis planning and execution time by approximately 40% compared with manual operation using RevMan and R scripts. Additionally, error rates in extracted effect sizes were comparable to those from human double entry. As rightly noted by the reviewer, a more comprehensive evaluation across multiple domains remains important. We are currently collecting additional performance metrics, including time-to-completion, extraction accuracy, inter-reviewer agreement and user trust, to further assess practical benefits.

Streamlining the systematic reviews and meta-analyses workflow

The SRMA workflow may be streamlined through the integration of AI-powered tools within a modular, pipeline-based architecture. In this setup, components such as literature screening, structured data extraction and narrative summarisation can be supported by LLMs. In contrast, analytical steps—such as effect size calculation, heterogeneity assessment and statistical visualisation—are handled by conventional statistical scripts or established meta-analysis software. The pipeline may be implemented using orchestration tools like LangChain, web scraping frameworks such as BeautifulSoup and data analysis libraries, with clear hand-offs between language understanding modules and computation engines.[55] It can continuously ingest new literature to maintain up-to-date synthesis. Containerisation tools like Docker encapsulate each module for reproducibility and portability. This hybrid approach leverages the reasoning capabilities

of LLMs while preserving analytical rigour, supporting scalable and reliable evidence generation.

Illustrative example: integration of the full LLM-enhanced SRMA framework

Our complete framework integrates multiple LLM components into a unified pipeline for the antidiabetic medication SRMA. The system periodically queries multiple databases using optimised search strategies for SGLT2i and GLP-1RA CV outcomes. New records are processed through a sequential workflow: the Screening Agent classifies abstracts; the Data Extractor uses comprehensive PICO templates for efficient field extraction; the Quality Assessor flags high-risk-of-bias studies using RoB 2.0 criteria; and the Analysis Engine automatically updates meta-analyses for heart failure hospitalisation, MACE, CV death and all-cause mortality outcomes. Results are presented through an interactive dashboard allowing real-time filtering by medication class, baseline CVD risk and study quality. The integrated system reduces evidence synthesis time from months to days while maintaining methodological rigour through strategic human checkpoints. This architecture demonstrates how multiple LLM components can function together as a cohesive system for living evidence synthesis in the rapidly evolving field of diabetes management.

DISCUSSION

Applying LLMs to SRMAs presents opportunities and challenges that warrant careful consideration. Our study contributes to an exploratory LLM-enhanced SRMA framework, intended as a starting point for academic discussion and practical experimentation, highlighting the potential to streamline various stages of SRMAs and optimise the process of SRAM such as literature screening, data extraction and synthesis. The application of LLMs offers substantial benefits, such as increasing research efficiency, reducing labour costs and the ability to process large volumes of information with consistency.

Collaboration models

In this framework, we proposed a synergistic cooperation pattern between LLMs and researchers, leveraging their strengths to enhance the efficiency and quality of SRMAs. With a vast knowledge base and advanced NLP capabilities, LLMs may serve as valuable consultants and assistants throughout the SRMA processes. LLMs, as consultants, may provide expert guidance and recommendations by analysing the research topic and identifying important knowledge gaps. With an extensive understanding of the research landscape, LLMs may suggest potential avenues for investigation that align with the researcher's objectives and have significant implications for practice or policy. LLMs, as assistants, may automate and streamline various tasks in the SRMA workflow. By applying NLP techniques, LLMs may efficiently process large volumes of literature, identify relevant studies and extract key information. This saved researchers valuable time and effort and reduced the risk of human error and bias.

Our framework views LLMs as complementary tools that augment and support the work of researchers, rather than as substitutes for human expertise. The human-in-the-loop system enables continuous refinement of collaboration between LLMs and researchers. By providing iterative feedback, researchers ensure that LLM-generated outputs are accurate, relevant and aligned with research objectives.[56] This feedback allows LLMs to fine-tune their behaviour to better align with researchers' needs and preferences, ultimately enhancing output quality and deepening their understanding of the research context.[57] However, a key challenge lies in balancing LLM reasoning capabilities with human oversight, as researchers may have limited access to the vast knowledge LLMs possess. To address this, our methodology emphasises the importance of transparent justifications for LLM outputs, such as supporting text segments, data sources and the rationale behind specific decisions. This transparency allows human reviewers to critically assess the reasoning process, ensuring alignment with ethical and scientific standards. Involving diverse human perspectives can further refine LLM outputs and help identify and mitigate potential biases.[58]

Potential bias

While LLMs exhibit the potential to reduce human biases in the SRMA processes, they can also introduce biases based on the data they are trained on[59] such as skewed data sets, the under-representation of certain demographics or the perpetuation of existing prejudices within the data. These biases may negatively impact the quality and reliability of the results. For example, position bias can lead LLMs to prioritise certain information based on its position within the text, potentially skewing literature selection or data extraction steps.[60] Furthermore, LLMs may exhibit confirmation bias, favouring outputs that align with their existing knowledge while overlooking conflicting information.[61] Data bias is also a concern, as LLMs trained on existing literature might reflect historical biases or fail to represent minority populations adequately.[62]

To address these issues, we propose a human-in-the-loop framework where researchers validate and refine LLM outputs throughout the SRMA process. Experts play a critical role in assessing the quality of results, providing context in prompts and correcting potential inaccuracies. This iterative feedback loop ensures that human expertise complements LLM capabilities, thereby maintaining the integrity of SRMAs. Moreover, selecting LLMs trained specifically on medical data and employing fine-tuning based on domain-specific data sets can help minimise these biases. Future research should explore advanced fine-tuning techniques and prompt engineering strategies to enhance the reliability and applicability of LLMs in SRMAs.

Hallucination reduction

The hallucination effect, where LLMs generate plausible but inaccurate information, can lead to potential inaccuracies in the SRMA processes.[63] While RAG technology can reduce hallucinations by grounding LLM outputs in external sources with appropriate chunk size and context length, it cannot fully eliminate them. This is particularly problematic when research questions or selection criteria fall outside the scope of the retrieved content, resulting in contextually plausible but inaccurate information. To address this, human-in-the-loop oversight is essential during critical phases such as research question formulation and study selection to ensure outputs are validated against source materials. Experts should identify contradictions based on their domain knowledge and trace sources for new information to verify the LLM's output.

Model expansion

Building on the LLM-enhanced SRMA framework, model expansion further optimises both consultant and assistant roles. Domain-specific LLMs (eg, PubMedBERT[64]) strengthen the consultant role by improving literature retrieval and research question formulation, while hybrid or agent-integrated models enhance the assistant role by automating subgroup analysis and bias assessment. Strategic selection of models streamlines RAG and agent-based workflows. General-purpose LLMs assist with summarisation and review writing, whereas specialised models optimise clinical outcomes analysis and effect size calculations, improving the overall quality and efficiency of SRMAs.

Our framework integrates a variety of LLM types, including general-purpose models (eg, GPT-4, Claude), domain-specific models (eg, PubMedBERT, BioMedLM), retrieval-augmented models (RAG-based variants) and agent-based orchestration systems. Each type supports different stages of the SRMA process, from broad text generation to specialised retrieval and structured data analysis.

To overcome challenges in handling specialised terminology, the framework employs two strategies: (1) providing few-shot examples or prompts to guide LLM outputs, and (2) fine-tuning with targeted domain-specific data to enhance contextual relevance and accuracy. Few-shot prompting is applied during literature screening, data extraction and bias assessment, while fine-tuning is primarily used for abstract screening and structured data extraction, where higher precision is critical. Although fine-tuning increases resource demands, it ensures that LLM outputs remain credible in specialised fields.

While RAG brings valuable benefits, such as access to up-to-date literature and dynamic subgroup analyses for precision medicine, its implementation presents challenges.[65] Ensuring the quality and relevance of retrieved information requires rigorous filtering protocols, including multistage retrieval validation, prompt consistency checks and expert adjudication. Integration of domain-specific knowledge, such as clinical guidelines, pharmacological references, outcome standards and epidemiological frameworks, further enhances the validity of findings. Nevertheless, the significant resource requirements of these systems raise concerns about scalability and sustainability.

Implementation challenges and risk mitigation

Despite the promise of LLMs, their real-world implementation faces several challenges. LLM-enhanced SRMAs require high-performance computational resources, such as secure cloud platforms and Graphics Processing Units (GPUs), to support efficient model inference and RAG workflows.[32] Researchers must also develop competencies in prompt engineering, critical evaluation of AI outputs and bias management to integrate LLMs effectively into evidence synthesis. To facilitate adoption, LLM systems should be modular and interoperable with established SRMA platforms such as Covidence, Rayyan and RevMan, enabling seamless workflow integration without major disruption.[41] In addition, the absence of clear regulatory standards and the need for updated institutional policies present barriers to the safe and ethical deployment of AI-assisted SRMAs, highlighting the necessity for proactive strategies to ensure responsible use.

To address novel risks introduced by LLM integration, the framework applies targeted mitigation strategies across each stage of the SRMA process. During literature retrieval, multisource validation and semantic similarity scoring minimise hallucination and irrelevant retrievals. Screening and study selection are conducted under human-in-the-loop oversight to ensure accurate application of eligibility criteria and reduce bias propagation. In data extraction, few-shot prompting combined with output cross-validation across multiple LLM instances helps to mitigate extraction errors and inconsistencies. For bias assessment, LLMs function as decision-support tools, while human reviewers retain final responsibility for judgments to prevent automation bias. During data analysis and synthesis, external expert reviews and detailed audit trails enhance transparency and detect potential model-driven misinterpretations. These safeguards collectively strengthen the reliability, transparency and ethical standards of LLM-enhanced SRMAs.

Ethical considerations

Integrating LLMs into SRMAs raises several important ethical issues that must be carefully managed.

First of all, although SRMAs generally rely on publicly available literature, researchers should ensure that no identifiable personal health information from preprints or case reports is inadvertently processed.[66] LLMs must operate within secure and compliant environments aligned with regulations such as General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA). Second, transparency is crucial. Researchers should disclose the extent of LLM involvement, including model versions, tasks assisted

and any prompt or fine-tuning strategies.[18] Such documentation enhances reproducibility and accountability. Third, accountability mechanisms are essential. Human reviewers must maintain final responsibility for SRMA outputs, verifying and validating LLM contributions. Maintaining audit trails and provenance records is recommended to trace and correct potential errors.[57] Finally, equitable access remains a challenge. Open-source LLMs and collaborative platforms are needed to ensure that researchers worldwide can leverage these technologies without disproportionate barriers.[65]

Future directions

This study proposes a conceptual framework for integrating LLMs into SRMAs, intended as a foundation for future empirical validation. While the framework is theoretically grounded and operationally detailed, its practical effectiveness requires systematic evaluation in real-world applications. In the future, developing LLM agents capable of autonomously supporting SRMAs will require close collaboration between researchers and computer scientists. Addressing current limitations, such as hallucinations and data privacy risks, will necessitate continuous technical refinement and ethical oversight. To promote widespread and responsible adoption, it will be essential to establish clear standards for the transparent reporting of AI-assisted SRMAs, ensuring reliability, reproducibility and methodological rigour. Future research should prioritise advancing hallucination mitigation techniques—such as dynamic retrieval augmentation and fine-tuned model calibration—and piloting the LLM-enhanced framework across diverse, high-impact fields, including oncology, infectious diseases and rare diseases. We acknowledge that the framework remains conceptual and unvalidated. A structured Delphi consensus process may offer a suitable path to assess its feasibility and completeness. These efforts will provide critical empirical validation, demonstrate practical value and refine the framework to better support the evolving needs of evidence synthesis.

CONCLUSIONS

This study presents an advanced SRMA framework enhanced by LLMs, incorporating RAG and agent-based methodologies. We demonstrate the application of LLMs at each stage of the SRMA processes and highlight its benefits compared with the traditional methodology. The functions of LLMs in SRMAs are categorised as consultants and assistants. It is argued that this innovative framework offers substantial value from both the technical and medical perspectives, as well as possible significant cost savings, with the potential to promote the transformation of SRMAs.

Acknowledgements The authors thank the reviewers and editors for their valuable comments and suggestions that improved this work. We thank UCL's Open access team for the open access support.

Contributors JS and JW proposed the research idea and contributed to conceptualisation. JS, ZL, DJ, SW and JW drafted the manuscript. All authors revised the manuscript, which was approved by all authors. JS is the guarantor.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no data sets generated and/or analysed for this study.

ORCID iD

Jinge Wu http://orcid.org/0000-0002-9873-0997

REFERENCES

1 Gopalakrishnan S, Ganeshkumar P. Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare. *J Family Med Prim Care* 2013;2:9–14.
2 Munn Z, Stern C, Aromataris E, *et al*. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Med Res Methodol* 2018;18:5.
3 Sedgwick P. Meta-analyses: heterogeneity and subgroup analysis. *BMJ* 2013;346:f4040.
4 Wang X-M, Zhang X-R, Li Z-H, *et al*. A brief introduction of meta-analyses in clinical practice and research. *J Gene Med* 2021;23:e3312.
5 Hansen C, Steinmetz H, Block J. How to conduct a meta-analysis in eight steps: a practical guide. *Manag Rev Q* 2022;72:1–19.
6 Borenstein M, *et al*. *Introduction to meta-analysis*. John Wiley & Sons, 2021.
7 Luo X, Chen F, Zhu D, *et al*. Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses. *J Med Internet Res* 2024;26:e56780.
8 Dai Z-Y, Wang F-Q, Shen C, *et al*. Accuracy of Large Language Models for Literature Screening in Thoracic Surgery: Diagnostic Study. *J Med Internet Res* 2025;27:e67488.
9 Scherbakov D, Hubig N, Jansari V, *et al*. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *J Am Med Inform Assoc* 2025;32:1071–86.
10 Cochrane methods IPD meta-analysis. Available: https://methods.cochrane.org/ipdma/frequently-asked-questions [Accessed 25 Apr 2024].
11 Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemp Clin Trials Commun* 2019;16:100443.
12 Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med* 2010;7:e1000326.
13 Wang Q, Feng Y, Huang J, *et al*. Large-scale generative simulation artificial intelligence: The next hotspot. *Innovation (Camb)* 2023;4:100516.
14 Hamel C, Hersi M, Kelly SE, *et al*. Guidance for using artificial intelligence for title and abstract screening while conducting knowledge syntheses. *BMC Med Res Methodol* 2021;21:285.

15 Zhang Y, Liang S, Feng Y, *et al*. Automation of literature screening using machine learning in medical evidence synthesis: a diagnostic test accuracy systematic review protocol. *Syst Rev* 2022;11:11.
16 Ouyang L, Wu J, Jiang X, *et al*. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst* 2022;35:27730–44.
17 Gemini Team Google. Gemini: a family of highly capable multimodal models. *arXiv* [Preprint] 2023.
18 Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *IOTCPS* 2023;3:121–54.
19 Ramkumar PN, Kunze KN, Haeberle HS, *et al*. Clinical and Research Medical Applications of Artificial Intelligence. *Arthroscopy* 2021;37:1694–7.
20 Wu H, Wang M, Wu J, *et al*. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *NPJ Digit Med* 2022;5:186.
21 van Dinter R, Tekinerdogan B, Catal C. Automation of systematic literature reviews: A systematic literature review. *Inf Softw Technol* 2021;136:106589.
22 Kohandel Gargari O, Mahmoudi MH, Hajisafarali M, *et al*. Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. *BMJ EBM* 2024;29:69–70.
23 Matsui K, Utsumi T, Aoki Y, *et al*. Large language model demonstrates human-comparable sensitivity in initial screening of systematic reviews: a semi-automated strategy using gpt-3.5. *SSRN* [Preprint].
24 Reason T, Benbow E, Langham J, *et al*. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. *Pharmacoecon Open* 2024;8:205–20.
25 Hasan B, Saadi S, Rajjoub NS, *et al*. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. *BMJ Evid Based Med* 2024;29:394–8.
26 Yan L, Martinez-Maldonado R, Gasevic D. Generative artificial intelligence in learning analytics: contextualising opportunities and challenges through the learning analytics cycle. LAK '24; Kyoto Japan, March 18, 2024 10.1145/3636555.3636856 Available: https://dl.acm.org/doi/proceedings/10.1145/3636555
27 Patil R, Gudivada V. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Appl Sci (Basel)* 2024;14:2074.
28 Minaee S, Mikolov T, Nikzad N, *et al*. Large language models: a survey. *arXiv* [Preprint] 2024.
29 Baldelli D, *et al*. TWOLAR: a two-step llm-augmented distillation method for passage reranking. European Conference on Information Retrieval; 2024
30 Gao L, Biderman S, Black S, *et al*. The pile: an 800gb dataset of diverse text for language modeling. *arXiv* [Preprint] 2020.
31 Tang Y-D, Dong E-D. LLMs in medicine: The need for advanced evaluation systems for disruptive technologies. *Innovation (Camb)* 2024;5:100622.
32 Lewis P, Perez E, Piktus A, *et al*. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv Neural Inf Process Syst* 2020;33:9459–74.
33 Gao Y, Xiong Y, Gao X, *et al*. Retrieval-augmented generation for large language models: a survey. *arXiv* [Preprint] 2023.
34 Kernan Freire S, Wang C, Foosherian M, *et al*. Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking. *Front Artif Intell* 2024;7:1293084.
35 Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *J Am Med Inform Assoc* 2025;32:605–15.
36 Han B, Susnjak T, Mathrani A. Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview. *Appl Sci (Basel)* 2024;14:9103.
37 Amugongo LM, Mascheroni P, Brooks SG, *et al*. Retrieval augmented generation for large language models in healthcare: a systematic review. *CSM* [Preprint] 2024.
38 Zhang Y, Li Y, Cui L, *et al*. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv* [Preprint] 2023.
39 Wang L, Ma C, Feng X, *et al*. A survey on large language model based autonomous agents. *Front Comput Sci* 2024;18:1–26.
40 Johri S, Jeong J, Tran BA, *et al*. Testing the limits of language models: a conversational framework for medical ai assessment. *medRxiv* [Preprint] 2023.
41 Cumpston MS, McKenzie JE, Welch VA, *et al*. Strengthening systematic reviews in public health: guidance in the *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd edition. *J Public Health (Bangkok)* 2022;44:e588–92.
42 Hammond MEH, Stehlik J, Drakos SG, *et al*. Bias in Medicine: Lessons Learned and Mitigation Strategies. *JACC Basic Transl Sci* 2021;6:78–85.
43 Korevaar DA, Salameh J-P, Vali Y, *et al*. Searching practices and inclusion of unpublished studies in systematic reviews of diagnostic accuracy. *Res Synth Methods* 2020;11:343–53.
44 Ahn E, Kang H. Introduction to systematic review and meta-analysis. *Korean J Anesthesiol* 2018;71:103–12.
45 Syriani E, David I, Kumar G. Screening articles for systematic reviews with ChatGPT. *J Comput Lang* 2024;80:101287.
46 Tan X, Shi S, Qiu X, *et al*. Self-criticism: aligning large language models with their understanding of helpfulness, honesty, and harmlessness. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Singapore, 2023 10.18653/v1/2023.emnlp-industry.62 Available: https://aclanthology.org/2023.emnlp-industry
47 Madaan A, Tandon N, Gupta P, *et al*. Self-refine: Iterative refinement with self-feedback. *Adv Neural Inf Process Syst* 2024;36.
48 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* 2023;11:e5339.
49 Borah R, Brown AW, Capers PL, *et al*. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 2017;7:e012545.
50 Higgins JPT, Altman DG, Gøtzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928.
51 Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. *Eur J Epidemiol* 2010;25:603–5.
52 Delgado-Chaves FM, Jennings MJ, Atalaia A, *et al*. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proc Natl Acad Sci USA* 2025;122:e2411962122.
53 Crowther M, Lim W, Crowther MA. Systematic review and meta-analysis methodology. *Blood* 2010;116:3140–6.
54 Bentaleb O, Belloum ASZ, Sebaa A, *et al*. Containerization technologies: taxonomies, applications and challenges. *J Supercomput* 2022;78:1144–81.
55 Asyrofi R, Dewi MR, Lutfhi MI, *et al*. Systematic literature review langchain proposed. 2023 International Electronics Symposium (IES); IEEE, Denpasar, Indonesia.
56 Cohn C, Snyder C, Montenegro J, *et al*. Towards a human-in-the-loop llm approach to collaborative discourse analysis. International Conference on Artificial Intelligence in Education; 2024
57 Amirizaniani M, Yao J, Lavergne A, *et al*. Developing a framework for auditing large language models using human-in-the-loop. *arXiv* [Preprint] 2024.
58 Yang J, Jin H, Tang R, *et al*. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Trans Knowl Discov Data* 2024;18:1–32.
59 Norori N, Hu Q, Aellen FM, *et al*. Addressing bias in big data and AI for health care: A call for open science. *Patterns (N Y)* 2021;2:100347.
60 Wang Z, Zhang H, Li X, *et al*. Eliminating position bias of language models: a mechanistic approach. *arXiv* [Preprint] 2024.
61 Gemalmaz MA, Yin M. Accounting for confirmation bias in crowdsourced label aggregation. Thirtieth International Joint Conference on Artificial Intelligence {IJCAI-21}; Montreal, Canada, 2021 10.24963/ijcai.2021/238 Available: https://www.ijcai.org/proceedings/2021
62 Yu Y, Zhuang Y, Zhang J, *et al*. Large language model as attributed training data generator: A tale of diversity and bias. *Adv Neural Inf Process Syst* 2024;36.
63 Ji Z, Yu T, Xu Y, *et al*. Towards mitigating llm hallucination via self reflection. Findings of the Association for Computational Linguistics; Singapore, 2023 10.18653/v1/2023.findings-emnlp.123 Available: https://aclanthology.org/2023.findings-emnlp
64 Cohn C. Bert efficacy on scientific and medical datasets: a systematic literature review. DePaul University, 2020.
65 Ntoutsi E, Fafalios P, Gadiraju U, *et al*. Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Min & Knowl 2020;10:e1356.
66 Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *NPJ Digit Med* 2024;7:183.

67 Adam GP, DeYoung J, Paul A, *et al*. *Literature search sandbox*: a large language model that generates search queries for systematic reviews. *JAMIA Open* 2024;7:ooae098.
68 Alshami A, Elsayed M, Ali E, *et al*. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems* 2023;11:351.
69 Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. *Syst Rev* 2024;13:219.
70 Guo E, Gupta M, Deng J, *et al*. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *J Med Internet Res* 2024;26:e48996.
71 Issaiy M, Ghanaati H, Kolahi S, *et al*. Methodological insights into ChatGPT's screening performance in systematic reviews. *BMC Med Res Methodol* 2024;24:78.
72 Cai X, Geng Y, Du Y, *et al*. Utilizing chatgpt to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. *Epidemiology* [Preprint] 2023.
73 Luo R, Sastimoglu Z, Faisal AI, *et al*. Evaluating the efficacy of large language models for systematic review and meta-analysis screening. *Health Informatics* [Preprint] 2024.
74 Khraisha Q, Put S, Kappenberg J, *et al*. 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;15:616–26.
75 Chen H, Jiang Z, Liu X, *et al*. Can large language models fully automate or partially assist paper selection in systematic reviews? *Br J Ophthalmol* 2025;109:962–6.
76 Oami T, Okada Y, Nakada T-A. Performance of a Large Language Model in Screening Citations. *JAMA Netw Open* 2024;7:e2420496.
77 Wang S, Scells H, Zhuang S, *et al*. Zero-shot generative large language models for systematic review screening automation. European Conference on Information Retrieval; 2024
78 Hasan B, Saadi S, Rajjoub NS, *et al*. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. *BMJ EBM* 2024;29:394–8.
79 Mahuli SA, Rai A, Mahuli AV, *et al*. Application ChatGPT in conducting systematic reviews and meta-analyses. *Br Dent J* 2023;235:90–2.
80 Kartchner D, Ramalingam S, Al-Hussaini I, *et al*. Zero-shot information extraction for clinical meta-analysis using large language models. The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks; Toronto, Canada, 2023 10.18653/v1/2023.bionlp-1.37 Available: https://aclanthology.org/2023.bionlp-1
81 Shah-Mohammadi F, Finkelstein J. Large language model-based architecture for automatic outcome data extraction to support meta-analysis. 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC); IEEE, Las Vegas, NV, USA.
82 Lam Hoai X-L, Simonart T. Comparing Meta-Analyses with ChatGPT in the Evaluation of the Effectiveness and Tolerance of Systemic Therapies in Moderate-to-Severe Plaque Psoriasis. *J Clin Med* 2023;12:5410.
83 Ji Z, Lee N, Frieske R, *et al*. Survey of Hallucination in Natural Language Generation. *ACM Comput Surv* 2023;55:1–38.
84 Farquhar S, Kossen J, Kuhn L, *et al*. Detecting hallucinations in large language models using semantic entropy. *Nature New Biol* 2024;630:625–30.
85 Majumder S, Dong L, Doudi F, *et al*. Exploring the capabilities and limitations of large language models in the electric energy sector. *Joule* 2024;8:1544–9.
86 Navigli R, Conia S, Ross B. Biases in Large Language Models: Origins, Inventory, and Discussion. *J Data and Information Quality* 2023;15:1–21.
87 Talukdar W, Biswas A. Improving large language model (llm) fidelity through context-aware grounding: a systematic approach to reliability and veracity. *arXiv* [Preprint] 2024.
88 Chang Y, Wang X, Wang J, *et al*. A Survey on Evaluation of Large Language Models. *ACM Trans Intell Syst Technol* 2024;15:1–45.
89 Wang B, Ping W, Xiao C, *et al*. Exploring the limits of domain-adaptive training for detoxifying large-scale language models. *Adv Neural Inf Process Syst* 2022;35:35811–24.
90 Xia Y, Zhang J, Jazdi N, *et al*. Incorporating large language models into production systems for enhanced task automation and flexibility. *arXiv* [Preprint] 2024.