informatics

MDPI

Review

An Agentic AI-Aided Review of Large Language Model Applications in Literature Reviews: A Seven-Layer Architecture

Emmanuel A. Merchán-Cruz [1,*], Ioseb Gabelaia [2], Shwe Soe [3], Irina Yatskiv [1] and Dmitry Pavlyuk [1]

[1] Engineering Faculty, Transport and Telecommunication Institute (TSI), LV-1019 Riga, Latvia; jackiva.i@tsi.lv (I.Y.); pavlyuk.d@tsi.lv (D.P.)
[2] Faculty of Business, Kauno Kolegija Higher Education Institution, LT-50468 Kaunas, Lithuania; iosebgabelaia@gmail.com
[3] Bristol Robotics Laboratory, University of the West of England, Bristol BS16 1QY, UK; shwe.soe@uwe.ac.uk
* Correspondence: merchan.e@tsi.lv

Abstract

The growing volume of scientific output and the pace of AI advancement create a dual challenge: traditional systematic reviews take twelve to eighteen months to complete, risking obsolescence before publication, while the technology needed to accelerate them is itself advancing faster than it can be reviewed. This work addresses both sides of that challenge simultaneously—presenting an agentic systematic literature review of AI-assisted reviewing, conducted using the same agentic pipeline it evaluates, across 122 primary studies published between 2022 and 2026. LLM performance proves to be sharply stage-specific: screening and structured extraction are tractable under human supervision, quality appraisal remains near-chance, and no published pipeline meets the bar for unsupervised end-to-end deployment. The review proposes a seven-layer reference architecture and identifies three prerequisite investments for production readiness: a shared benchmark, validated risk-of-bias performance, and a mandatory AI reporting standard.

Keywords: systematic review; large language models; agentic AI; PRISMA 2020; screening automation; data extraction; risk of bias; reporting transparency; human in the loop

1. Introduction

1.1. The Obsolescence of the Twelve-Month Systematic Review

The systematic review has long served as the epistemic foundation of evidence-based practice, providing the highest-quality signal available to clinicians, policymakers, and researchers by exhaustively retrieving, appraising, and synthesising all available evidence on a defined question. Yet the infrastructure on which that signal rests is under acute strain. The primary scientific literature is expanding at a rate that fundamentally outpaces traditional review methodology: the scientific corpus doubles every 9 years [1], with more than 2.5 million papers published annually across approximately 30,000 accredited journals worldwide [2]. The growth is not confined to biomedicine; in applied linguistics, for instance, relevant search returns more than doubled between 2009 and 2024 [3].

Consequently, traditional manual systematic literature reviews (SLRs) are labour-intensive, error-prone, and increasingly unsustainable, demanding substantial investments of time, resources, and financial capital [4]. A typical review requires an average of 67.3 weeks from inception to completion, with the screening phase alone consuming hundreds of person-hours: one comparative case study estimated 564.4 h to complete

Check for updates

Academic Editor: Yang Shi

Received: 12 June 2026
Revised: 10 September 2026
Accepted: 24 September 2026
Published: 1 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

both the title/abstract and full-text screening phases for an initial pool of 14,439 records, while another team required 96–110 h to screen just 3258 [1,5]. Reviewers working at this scale sustain significant fatigue, a recognised source of systematic error at the eligibility screening stage [6].

The scalability crisis is not uniform across the pipeline. Three large-scale methodological reviews map the AI development landscape and converge on a structural asymmetry in automation effort: Köhler et al. [7] (54 studies) finds screening most covered (79.6%) and quality appraisal least (18.5%); Scherbakov et al. [8] (172 citations) places search (34.9%), extraction (31.4%), and title/abstract screening (25.0%) ahead of full-text screening (8.1%), quality appraisal (7.0%), and drafting (12.8%); Lieberum et al. [9] (37 studies) reports 41%/38%/30% for search/screening/extraction against 14%/5.4% for appraisal/reporting. All three converge: early-stage tasks dominate development effort, while quality appraisal and manuscript reporting remain neglected.

Figure 1 illustrates this asymmetry across all three mapping reviews: screening consistently receives the greatest investment while quality appraisal remains the most underserved stage (7.0–18.5% coverage across the three reviews).

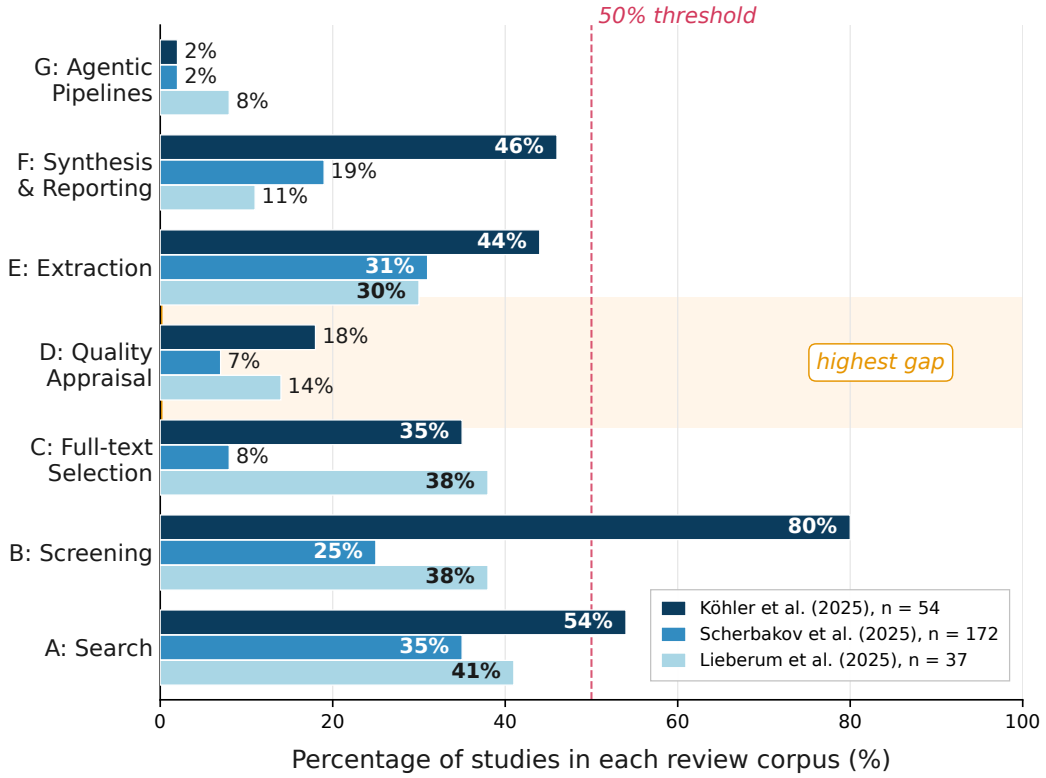


Figure 1. Stage asymmetry in AI-assisted systematic reviewing: the proportion of corpus studies applying AI assistance at each PRISMA-defined stage, across three large-scale methodological mapping reviews [7–9].

This uneven penetration is consequential. At the screening stage—where AI development effort has concentrated most heavily—LLMs achieve high sensitivity but at the cost of low precision, generating substantial false positives that require human adjudication. Performance metrics span a wide continuum across model configurations and deployment contexts, as detailed in Section 4.2. The challenge is therefore not whether LLMs can screen, but under what conditions and configurations they do so safely.

1.2. From AI-Assisted to Agentic Systematic Reviews

The studies above largely exemplify *AI-assisted* reviewing: a human delegates one defined sub-task to a language model, evaluates the output, and retains decision authority. The more ambitious *agentic* paradigm has an AI system plan and execute multiple stages sequentially, passing outputs between stages and invoking tools without continuous human intervention. Köhler et al. [7] employ a four-level taxonomy—Human-Centric (29.6%), Human-in-the-Loop (40.7%), AI-Driven with Human Validation (27.8%), Autonomous AI (1.9%)—showing true autonomous execution is exceedingly rare. Luo et al. [4] contrast conversational chatbots with LLM-driven autonomous agents capable of screening, extraction, and composition, while cautioning that strategic planning and complex problem solving remain open challenges. Early multi-stage demonstrations are uneven: Sollini et al. [10] spans search through synthesis with universal failure at PRISMA-compliant drafting; Purewal et al. [11] reports near-perfect pooling accuracy; Roy et al. [12] tackles query formulation and retrieval. Performance varies sharply by stage within every multi-step system evaluated (Section 7).

The corpus converges emphatically on the necessity of human-in-the-loop configurations: hallucination and the unreliable execution of nuanced tasks remain pervasive [13–16], compounded by processing constraints such as context-window limits [13,14,16] and prohibitive API document-upload costs [15]. Pooled meta-analytic evidence shows overall sensitivity and specificity approaching 90% [17], with highly variable workload reductions ranging from 33% to 93% for a second human reviewer in hybrid title and abstract screening configurations [18] (Sections 4.2–7). A further deployment barrier is the absence of mandatory AI reporting standards in PRISMA 2020 [19]: temperature is disclosed in $<$2% of papers and verbatim prompts in $<$10% (Section 6.3), leaving large fractions of the evidence base non-reproducible.

1.3. Research Questions and Contributions

Against this backdrop, four research questions motivate this systematic review:

RQ1 What capabilities have LLMs demonstrated at each of the seven principal stages of a PRISMA-defined systematic review (search-query generation, screening, data extraction, quality appraisal, evidence synthesis, reporting, and end-to-end agentic execution), and what are the documented performance metrics and failure modes at each stage?

RQ2 Which LLM architectures, prompt engineering strategies, and tool configurations produce the highest task performance at each stage, and do those findings generalise across domains and review types?

RQ3 What evaluation frameworks, benchmarks, and reporting standards currently govern AI-assisted and agentic systematic review, and how adequate are they?

RQ4 What are the necessary technical, methodological, and regulatory preconditions for a validated, general-purpose agentic systematic review system, and what is the evidence-based path to realising them?

This review makes five contributions to the literature.

First, it provides the first systematic stage-by-stage mapping of LLM applications to all seven PRISMA-defined SLR stages across a corpus of 122 primary studies published between 2022 and 2026. The mapping reveals a pronounced stage asymmetry that, as noted in Section 1.1, is consistent across all three large-scale methodological reviews: early-stage tasks are most frequently automated, while quality appraisal and manuscript reporting remain comparatively neglected.

Second, it produces a taxonomic classification of the 122 corpus papers by automation depth—single-stage AI assistance, multi-stage AI coordination, and end-to-end agentic

execution—anchored by Köhler et al.'s [7] four-level taxonomy—and demonstrating that fully autonomous agentic execution currently represents a small fraction of the literature, which shares no common evaluation standard.

Third, it conducts the first systematic comparison of LLM performance across SLR stages using a common analytic lens, allowing head-to-head comparison of tool ecosystems (GPT-4/GPT-4o, Claude, Gemini, DeepSeek, Llama, and open-source fine-tuned models) and identifying which model families dominate which stages.

Fourth, drawing on the convergent gaps identified across all four research questions, the review derives a structured research agenda specifying the benchmarking infrastructure, reporting standards, and architectural requirements that must be established before agentic systematic review can be considered fit for unsupervised deployment.

Fifth, and most distinctively, the review does not merely document the gap—it partially fills it. The pipeline used to conduct this review is itself an operational instantiation of the seven-layer reference architecture proposed in Section 10.2: an MCP-connected agentic workflow spanning automated multi-database search, multi-stage hybrid screening, retrieval-augmented synthesis grounded exclusively in an uploaded paper corpus, and a two-stage citation audit applied to every in-text claim, with human author oversight and sign-off at every stage. The methodology therefore constitutes the first documented attempt to execute a PRISMA-compliant systematic review using a fully specified agentic architecture—making the paper simultaneously a study of agentic reviewing and an example of it.

1.4. Roadmap

The remainder of this paper is organised as follows. Section 2 describes the systematic search strategy and review protocol. Sections 3–7 present corpus characterisation and thematic results by SLR stage, culminating in a cross-stage synthesis matrix (Section 8). Section 9 interprets findings against the four research questions, and Section 10 states conclusions, proposes a seven-layer reference architecture, and documents how the methodology of this review constitutes an early operational instantiation of that architecture.

2. Methods

2.1. Protocol and Registration

This systematic literature review (SLR) was conducted and reported following PRISMA 2020 guidance [19]. The review protocol was not pre-registered on PROSPERO; this constitutes a limitation acknowledged in Section 9.2, arising from the exploratory and methodologically heterogeneous scope of the corpus, which precluded the specification of a sufficiently precise a priori protocol before the search was executed. No eligibility changes were made post-search. To mitigate the absence of prospective registration, the review protocol as executed—eligibility criteria, information sources, and the selection process—is provided as a timestamped, publicly citable document (Zenodo, DOI 10.5281/zenodo.22691555). The study uses narrative thematic synthesis following Thomas and Harden [20], which was selected because the corpus spans multiple methodological designs and LLM architectures that cannot be meaningfully pooled through meta-analytic statistics.

2.2. Eligibility Criteria

Studies were eligible for inclusion if they satisfied all five inclusion criteria (IC1–IC5) and none of the five exclusion criteria (EC1–EC5), as specified in Table 1. Criteria were operationalised prior to screening and applied consistently throughout.

The central discriminating criterion was IC4: the paper must explicitly apply, evaluate, or propose use of an LLM or generative AI tool (e.g., ChatGPT, GPT-4, Claude, Gemini, Llama, Elicit, Rayyan AI) in one or more stages of a systematic review, scoping review, or evidence synthesis. Two principal false-positive categories were excluded: papers reviewing LLMs as a subject rather than as a methodological tool, and papers using only traditional ML/NLP (BERT, SVM) without a generative component (EC1).

Table 1. Inclusion and exclusion criteria.

Type	Code	Label	Criterion
Inclusion	IC1	Study type	Peer-reviewed journal article, conference paper (≥4 pages), or SR protocol
Inclusion	IC2	Language	English language only
Inclusion	IC3	Date	Published January 2022–March 2026
Inclusion	IC4	AI content	Explicitly applies, evaluates, or proposes use of an LLM or GenAI tool in ≥1 stage of an SR, scoping review, or evidence synthesis
Inclusion	IC5	Empirical	Reports ≥1 empirical outcome (performance metric, efficiency measure, error rate, usability score, or qualitative assessment)
Exclusion	EC1	Traditional ML	Uses only traditional ML/NLP (BERT, SVM, random forest) without a generative AI component
Exclusion	EC2	Bibliometrics	Focused solely on reference management or bibliometrics without AI review tasks
Exclusion	EC3	Study type	Editorial, opinion, letter, book chapter, thesis, or non-peer-reviewed
Exclusion	EC4	Language	Not in English
Exclusion	EC5	Peripheral AI	AI tools mentioned only peripherally (e.g., grammar checking only, not SR tasks)

2.3. Information Sources and Search Strategy

Three bibliographic databases were searched: Scopus (multidisciplinary), PubMed (health sciences), and IEEE Xplore (computing/engineering). These databases were selected to balance topical coverage with search reproducibility: Scopus offers broad multidisciplinary indexing that substantially overlaps Web of Science and indexes a large proportion of IEEE and ACM venues, PubMed covers the health and biomedical literature, and IEEE Xplore ensures complete coverage of core computing and engineering venues (coverage limitations are discussed in Section 9.2). No grey literature or preprint servers were included (Section 9.2). The search was executed in March 2026, covering January 2022–March 2026 (lower bound coinciding with the period in which general-purpose LLMs first became widely accessible to the research community [21]).

Search strings were constructed around three orthogonal concept blocks combined with Boolean AND operators: (i) *evidence synthesis process terms*—"systematic review", "scoping review", "evidence synthesis", "literature review"; (ii) *AI tool terms*—"large language model", "LLM", "generative AI", "ChatGPT", "GPT-4", "GPT-3", "Claude", "Gemini", "Llama", "Elicit", "Consensus", "Rayyan", "foundation model", "instruction-tuned"; (iii) *SR task terms*—"screen*", "extract*", "apprais*", "search strateg*", "synthesis", "automat*", "evidence table*". Database-specific field tags (TITLE-ABS-KEY for Scopus; [tiab] for PubMed; document-type and language filters for all three) were applied as appropriate. The full Boolean search strings for all three databases are provided in the Supplementary Materials (Appendix S1) and openly deposited in a citable repository (Zenodo, DOI 10.5281/zenodo.22691555).

The initial search yielded 1786 records: Scopus ($n = 1174$), PubMed ($n = 561$), and IEEE Xplore ($n = 51$). Records were exported in RIS format and processed programmatically. Cross-database deduplication was performed using DOI as the primary matching key (present in 99.0% of records) and normalised title + year as the secondary key; 22 duplicate records were identified and removed, with the Scopus version retained where a record appeared in multiple databases.

2.4. Software Tools and AI-Assisted Review Pipeline

AI assistance was integrated at three pipeline stages and documented in accordance with emerging reporting standards for AI-assisted reviews [22]: (1) **Pre-screening**: Claude (Anthropic, `claude-sonnet-4-6`) operating via the Model Context Protocol (MCP) executed regex-based IC4-signal detection on the full RIS corpus (Section 2.5); all classifications were reviewed by the human author team. (2) **Synthesis**: Google NotebookLM served as a constrained retrieval-augmented generation (RAG) environment; the 122 included papers were partitioned across six thematic notebooks (T1–T6; Section 2.7), with each queried with APA 7 citation-preamble prompts (Appendix A). (3) **Co-drafting**: Claude (`claude-sonnet-4-6`) via MCP assisted with citation key mapping and manuscript drafting; all text was reviewed and approved by the human author team. The citation audit log is summarised in Appendix B. The underlying architecture is described in Section 2.10.

Figure 2 provides an overview of the three-phase AI integration pipeline and the human oversight gates that governed all AI-assisted decisions.

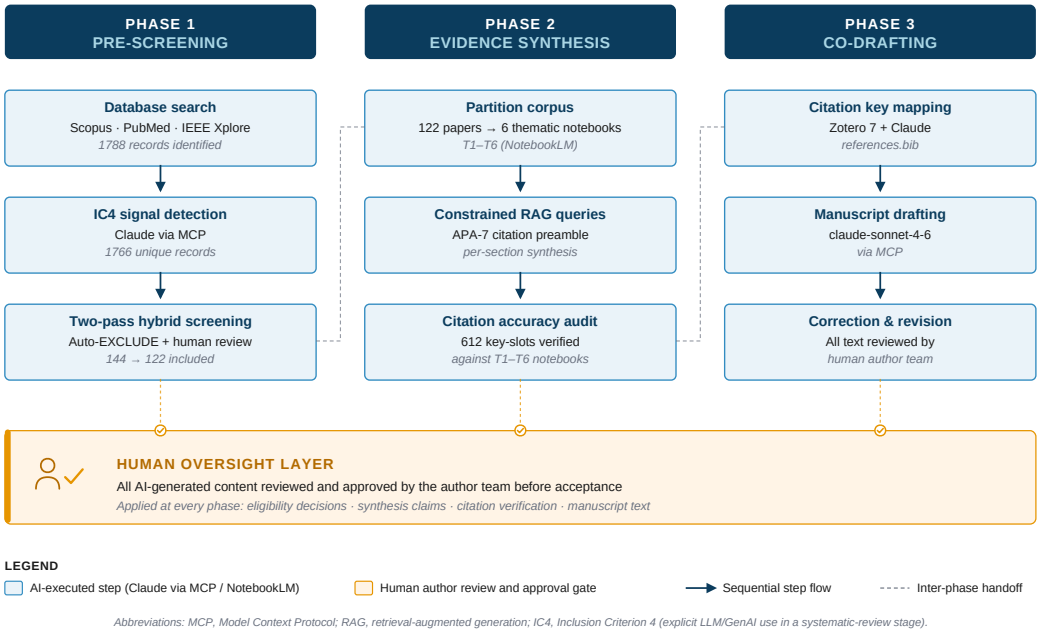


Figure 2. AI-assisted review pipeline used in this study. The three integration phases (pre-screening, evidence synthesis, co-drafting) and the corresponding tools are shown alongside the mandatory human oversight gates that validate all AI-generated content before acceptance.

2.5. Study Selection

Study selection was conducted in two sequential stages following the PRISMA 2020 screening protocol.

Stage 1 (title-and-abstract screening) was performed on the 1764 remaining records using a two-pass hybrid approach. In the first pass, an algorithmic pre-screener applied a suite of regular-expression signals designed to detect IC4-eligible content (LLM/GenAI tool applied to an SR stage) and to flag known false-positive patterns (e.g., the verb "elicit", SLR-topic papers mentioning AI in clinical domains). Records meeting no positive IC4 signal and no ambiguity trigger were classified as EXCLUDE; records matching IC4 signals were classified as INCLUDE or UNCERTAIN. In the second pass, the author team reviewed all 10 UNCERTAIN records with the decision to include all 10; in addition, a random 10% sample of auto-EXCLUDE records was manually reviewed to verify recall, applying the IC/EC criteria to title and abstract text. This recall check re-screened a random 10% sample of the auto-EXCLUDE records ($n = 161$) against IC1–IC5; no eligible record was identified

as missed, giving an estimated false-negative rate of 0% (one-sided 95% upper bound ≈ 1.9% by the rule of three, 3/161). Final inclusion/exclusion decisions at Stage 1 rested entirely with the human reviewers, rendering 144 papers.

Stage 2 (full-text eligibility review) was conducted for the 144 records surviving Stage 1. Full-text PDFs were retrieved via institutional access to verify IC4 and IC5; 12 of the 144 records could not be retrieved (institutional and open-access routes exhausted) and were excluded, leaving 132 reports assessed for eligibility. A further 10 reports were excluded at full-text review for failing IC4 or IC5, yielding a final corpus of 122 papers (PRISMA 2020 flow diagram, Section 3.1). The full list of full-text exclusions with reason codes is provided in the Supplementary Materials (Table S2).

The use of AI assistance in the first-pass pre-screening is transparently acknowledged as a methodological feature of this review—and is itself an example of the practice under study. This creates a limited form of circularity (the review uses an LLM to partly conduct an LLM-in-SR review) that the authors treat as a strength from a reflexivity standpoint and a limitation from an independence standpoint; both aspects are discussed in Section 9.2.

Stage 1 produced the following counts by database (Table 2):

Table 2. Title-and-abstract screening results by database.

Database	Screened	INCLUDE	EXCLUDE	UNCERTAIN	Include Rate
Scopus	1174	116	1056	2	9.9%
PubMed	561	25	530	6	4.4%
IEEE Xplore	51	3	46	2	5.9%
Total retrieved	1786	144	1632	10	8.1%
After deduplication †	1764	144	1610	10	8.2%

† A total of 22 cross-database duplicate records were removed prior to screening (19 PubMed–Scopus; 3 IEEE–Xplore–Scopus); the Scopus version was retained in all cases. All counts in the "After deduplication" row reflect the unique 1764 records that were actually screened.

The dominant exclusion reason was that IC4 was not met (AI or LLM is the subject of a review, not a tool used in its conduct), which accounted for 99.3% of all excluded records ($n = 1599$). Exclusions under EC1 (traditional ML/NLP without generative component, $n = 4$) and EC3 (non-peer-reviewed, $n = 7$) accounted for the remainder.

2.6. Data Extraction

Data extraction was conducted through an agentic workflow in which Claude (Anthropic, `claude-sonnet-4-6`, San Francisco, CA, USA), operating via the Model Context Protocol (MCP), queried each paper's dedicated NotebookLM notebook (T1–T6; Section 2.4) using a pre-agreed query plan aligned to a defined study structure specified by the authors. For each included paper the agent extracted: bibliographic metadata, AI tool(s) and model version, SR stage(s) targeted, study design type, performance metrics reported, evaluation corpus size, subject domain, presence of a human comparator (yes/partial/no), and principal finding. All extracted fields were reviewed and approved by the human author team before acceptance; any discrepancy flagged by the agent was resolved through direct inspection of the source paper.

2.7. Thematic Synthesis

Thematic synthesis followed the three-stage framework of Thomas and Harden [20]: (1) inductive coding of each paper to identify SR stage(s) targeted and AI contribution type; (2) grouping into thematic clusters; and (3) development of analytical themes mapped to RQ1–RQ4. Six thematic clusters emerged (Table 3). T5 is heterogeneous by design, grouping multi-stage automation papers, critical appraisals, and governance frameworks.

A cross-cutting *Human–AI Teaming* (HAT) flag was applied to papers reporting quantitative LLM–human agreement ($n = 54$, 44.3%).

Table 3. Thematic taxonomy applied to the 122 included papers.

Code	Theme	Definition	*n*	%
T1	Screening & Study Selection	LLMs/GenAI applied to title/abstract screening, citation triage, or full-text study selection	35	28.7
T2	Search Strategy & Query Generation	LLMs used to construct Boolean search strings, PICO queries, or MeSH terms	8	6.6
T3	Data Extraction & Evidence Synthesis	LLMs used to extract PICO elements, populate evidence tables, or generate synthesis	13	10.7
T4	Quality Assessment & Methodological Appraisal	LLMs applied to RoB 2, ROBINS-I, AMSTAR-2, PRISMA adherence, or CONSORT	7	5.7
T5	Integrated Pipelines, Frameworks & Critical Overviews	Multi-stage automation frameworks, cross-task benchmarks, or conceptual frameworks for responsible AI in evidence synthesis	36	29.5
T6	Domain-Specific AI-Conducted Reviews	LLM used as primary agent to execute or substantially assist an SR in a specific subject domain	23	18.9
Total			122	100.0

Papers were assigned to their primary thematic cluster based on the stage at which LLM assistance was the principal methodological contribution. Where a paper addressed two or more SR stages approximately equally, it was assigned to T5. This single-assignment approach was chosen for clarity of presentation; thematic overlap is discussed qualitatively in Section 9.

2.8. Notation and Nomenclature

This review uses four distinct labelling systems. To prevent ambiguity—in particular, any confusion between automation levels and architecture layers—they are summarised in Table 4. The prefixes are kept deliberately distinct: thematic clusters use **T1–T6**, review stages use **Stage A–G**, automation levels use **L0–L3**, and the reference-architecture layers are written out in full as **Layer 1–7** (never abbreviated to "L"), so that "L" always denotes an automation level.

Table 4. Notation used throughout this review.

Labelling System	Notation	Meaning and Where Defined
Thematic clusters	T1–T6	Six corpus themes grouped by SR task (Table 3, Section 2.7).
Review stages	Stage A–G	PRISMA-ordered stages (canonical, per Figure 1): A Search, B Screening, C Full-text selection, D Quality appraisal, E Extraction, F Synthesis & reporting, G Agentic pipelines (Section 3.3).
Automation levels	L0–L3	Degree of AI autonomy, from L0 (manual) to L3 (fully autonomous) (Section 7).
Architecture layers	Layer 1–7	The proposed seven-layer reference architecture: 1 Protocol, 2 Search, 3 Screening, 4 Extraction, 5 Quality appraisal, 6 Synthesis, 7 Transparency (Section 10.2).

2.9. Quality Assessment of Included Studies

Given the heterogeneity of study designs, formal meta-analytic pooling was not appropriate. A modified MMAT v2018 [23] was applied to each paper, primarily under the *quantitative descriptive* and *mixed methods* categories. MMAT scores were not used as exclusion thresholds but rather to contextualise confidence in individual findings reported in Sections 3–9; the full per-study MMAT ratings are provided in the Supplementary Materials and openly deposited (Zenodo, DOI 10.5281/zenodo.22691555).

2.10. Proposed Agentic AI-Assisted Review Architecture

The review was implemented as a tool-augmented, agentic architecture in which Claude (Anthropic) operates via the Model Context Protocol (MCP)—an interoperability framework enabling direct LLM interaction with external data sources, file systems, and analytical tools. Unlike conventional prompt-based workflows, the MCP-enabled agent maintains persistent state across interactions and invokes tools dynamically, executing multi-step tasks (query formulation, pre-screening, classification, data extraction) without manual mediation between steps.

The architecture, illustrated in Figure 3, is organised around a central MCP-enabled orchestrating agent—Cowork—which manages Query Execution, Screening Logic, Classification, and Task Orchestration across all review stages. Bibliographic data sources (Scopus, PubMed, and IEEE Xplore) feed into the agent on the left, providing the raw retrieval corpus. On the right, the agent produces three classes of structured output: a Screened Corpus, an Extracted Dataset, and a Manuscript Draft. Below the agent, three Tool-Augmented Processing Modules extend its capabilities: Zotero handles Reference Management, Deduplication, and Metadata Storage; VOSviewer 1.6.20 supports Bibliometric Mapping and Cluster Detection; and NotebookLM provides a Full-Text Corpus environment, Thematic Synthesis, and an Extended RAG interface for grounded generation. Human-in-the-Loop Validation and Control operates above the entire system, with bidirectional Oversight and Feedback arrows linking the human author team to the agent at every stage; all decisions, extraction outputs, and synthesised content are validated by the human author team before acceptance, supporting PRISMA 2020-guided reporting [19].

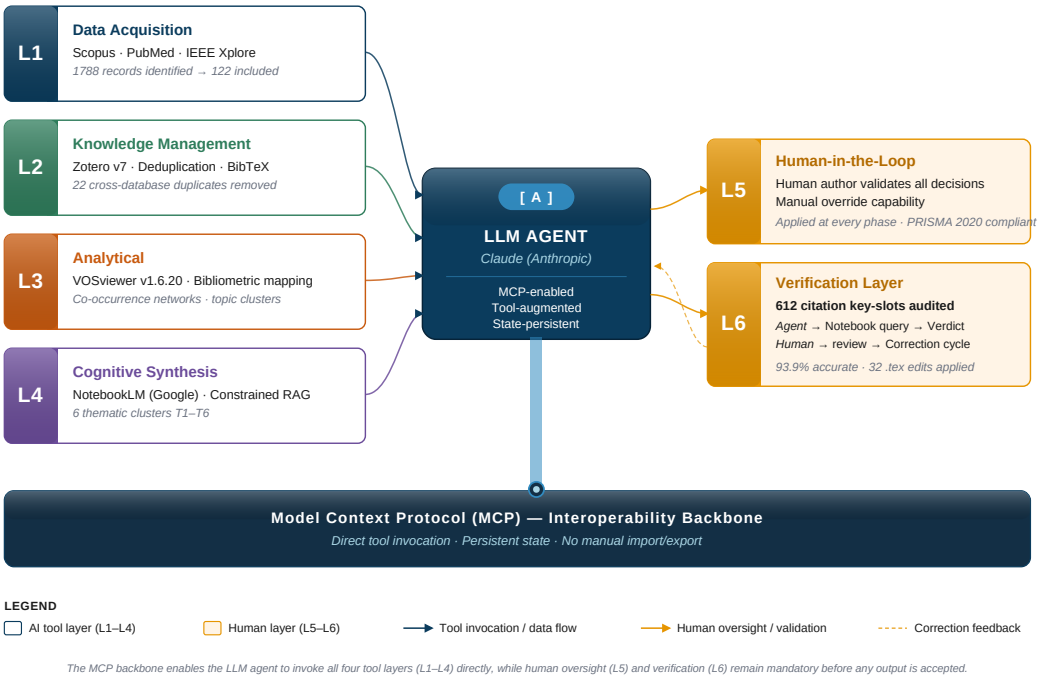


Figure 3. MCP-enabled agentic review architecture. Cowork (central hub) orchestrates Query Execution, Screening Logic, Classification, and Task Orchestration via the Model Context Protocol. Bibliographic data sources (Scopus, PubMed, IEEE Xplore) supply the retrieval corpus (left); Tool-Augmented Processing Modules—Zotero, VOSviewer, NotebookLM—extend agent capabilities (below); structured outputs (Screened Corpus, Extracted Dataset, Manuscript Draft) are produced on the right; Human-in-the-Loop Validation and Control provides bidirectional oversight above the entire system.

2.11. Transparency and Artificial Intelligence Disclosure

This review was partially executed using large language model assistance. Full transparency is provided in accordance with PRISMA 2020 item 24 and the emerging consensus on minimum reporting standards for AI-assisted reviews, as characterised by the included studies themselves.

Tools and models used. Claude Sonnet 4.6 (Anthropic, `claude-sonnet-4-6`; accessed March–April 2026) served as the primary synthesis engine throughout the project. Google NotebookLM (accessed April 2026) was used as the corpus-management and evidence-retrieval platform: all 122 included papers were uploaded to six thematic notebooks (T1–T6 as documented in the project protocol), and each synthesis query was submitted to the relevant notebook with a mandatory full-paper reading preamble requiring the system to read introduction, methods, results, discussion, and conclusions of every source before answering. Claude in Chrome (Anthropic; accessed April 2026) provided browser automation enabling systematic multi-tab query execution across all six notebooks in parallel. No AI tool was used in the final study selection process: title/abstract and full-text inclusion/exclusion decisions were made by the human author team.

Stages automated and stages retained by humans. AI assistance was applied to: (a) drafting synthesis paragraphs from structured notebook query outputs; (b) populating data extraction tables from notebook-confirmed evidence; (c) generating initial section drafts subsequently revised by human authors against the primary studies; (d) executing complete Phase 3 rewrites of all results and discussion sections from live NotebookLM synthesis queries, each grounded exclusively in the uploaded paper corpus; and (e) a systematic Phase 4 citation accuracy audit in which every in-text citation key-slot across all sections was individually verified by a live NotebookLM query before acceptance. All interpretive claims and the overall argumentative framing of the review were produced and approved by the human author team. Study eligibility decisions, risk-of-bias judgements, and final conclusions were made exclusively by human reviewers.

Prompt disclosure. A representative selection of the NotebookLM query prompts used during synthesis is provided in Appendix A, including the mandatory full-paper reading preamble that prefaced each query. Phase 4 citation-audit query–response pairs are summarised in Appendix B and the accompanying `Citation_Audit_Log.xlsx` workbook.

Model parameters. Claude Sonnet 4.6 was accessed via the Anthropic API at default parameters (temperature not explicitly set; Anthropic default-applies at inference time). NotebookLM queries were submitted via the standard web interface; NotebookLM does not expose temperature or top-p parameters to end users. Only two studies in our corpus disclose inference parameters: Tripathi et al. [5] reports temperature = 0.1 and top-p = 0.95; Yun et al. [24] reports top-p = 0.7 and max_length = 2048. Additionally, Soprano et al. [25] explicitly mapped the stages of their AI-assisted pipeline to PRISMA 2020 items 4–10b, 13a, 16a, 20a–d, 21, and 22. While several corpus studies provided standard PRISMA 2020 checklists in their supplementary materials, the present review exceeds the median corpus practice on all three transparency dimensions: PRISMA item mapping, prompt disclosure, and model parameter reporting.

2.12. Citation Audit and Verification Protocol

A two-stage citation audit protocol was implemented as an independent verification layer operating after synthesis and before finalisation, addressing the provenance-tracking limitation inherent in generative AI systems [24,26].

In **Stage 1 (agent-based verification)**, the LLM agent submitted each citation's surrounding paragraph to the source notebook (T1–T6), recording a structured verdict: *Supported* (claim fully stated in cited paper), *Partially supported* (over-attribution detected),

Unsupported (claim contradicts or is absent from the cited paper), or *Cannot verify* (paper not loaded in the queried notebook).

In **Stage 2 (human validation)**, the human author team recorded an independent determination for every key-slot in the audit log's Status field—not only those the agent flagged. The 535 slots the agent judged *Accurate* were each confirmed by a human reviewer against the cited paper (Status: *Verified*); the 35 *Partially Accurate* and 20 *Needs Correction* slots were re-read and resolved (Status: *Corrected* or *Accepted*); and the 22 *Cannot Verify* slots—citations to complementary references outside the 122-paper corpus that the authors introduced for additional context—were checked against those complementary sources and accepted. Every citation key-slot therefore passed through independent human verification, so the reported accuracy is a human-validated result rather than a self-assessment by the agent.

The two-stage protocol ran to completion across all sections of the manuscript and Appendix C. All 612 citation key-slots were audited and human-validated; the 36 slots requiring correction were revised in the source, and the human determination for every slot is recorded in the Status field of the audit log.

The two-stage workflow is summarised in Figure 4. All audit records are maintained in *Citation_Audit_Log.xlsx* (one row per citation key-slot; schema in Table 5). Per-section aggregate results are provided in Supplementary Table S7.

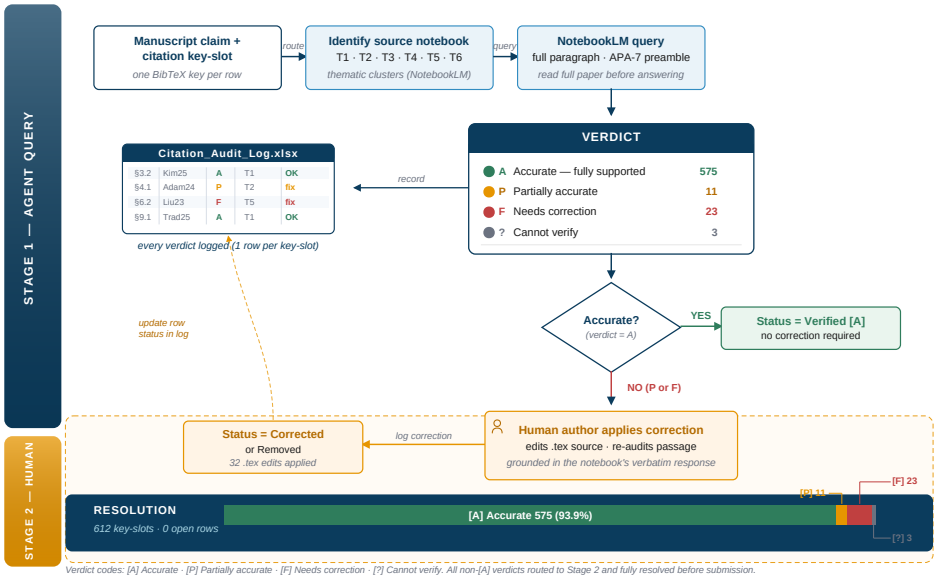


Figure 4. Two -stage citation verification workflow. Stage 1 (agent-based): the LLM agent queries each paper's source notebook (T1–T6) and records a structured verdict per citation key-slot. Stage 2 (human validation): the author team records an independent determination for every key-slot, applies corrections to the `.tex` source, and accepts the resolved key-slot. All 612 citation key-slots completed both stages.

Table 5. Citation audit log column schema (Supplementary Materials, *Citation_Audit_Log.xlsx*, Sheet 1).

Col	Header	Content
A	Section	Sections 1–10
B	Subsection/Para	LaTeX heading and paragraph number
C	Claim (Verbatim)	Full paragraph text in which this key appears
D	Citation Key	Single BibTeX key; one key per row
E	Verdict	Accurate/Partially Accurate/Needs Correction/Cannot Verify
F	Issue Identified	Verbatim notebook response if verdict ≠ Supported
G	Recommended Correction	Revised `.tex` text or "None required"
H	Verification Source	Notebook label (T1–T6) + date
I	Status	Verified/Corrected/Accepted

The Present Review as a Proof-of-Concept Articulation

Taken together, the methodological stages described in Sections 2.1–2.12 constitute an early articulation of the seven-layer reference architecture proposed in Section 10.2. The mapping is direct: the structured PICO/PCC eligibility criteria injected into every downstream agent prompt correspond to Layer 1 (Protocol); the LLM-assisted Boolean query generation with mandatory human sign-off to Layer 2 (Search); the two-pass hybrid screener to Layer 3 (Screening); the MCP-connected Claude extraction workflow to Layer 4 (Extraction); the fully human quality appraisal and study classification to Layer 5 (Appraisal); the NotebookLM retrieval-augmented synthesis with human author review to Layer 6 (Synthesis); and the two-stage citation audit, verbatim audit-evidence files, and the AI-disclosure subsection (Section 2.11) to Layer 7 (Transparency). The present review therefore serves not only as an analysis of agentic systematic review in the literature but as a worked instantiation of the architecture it proposes—one conducted under continuous human oversight and subject to the same verification standards recommended for future production deployments.

2.13. Evaluation Metrics

All metrics reported in Sections 3–8 are those actually used in the included papers; Table 6 summarises the principal metrics by SR stage.

Table 6. Principal evaluation metrics by SR stage (corpus-derived).

Stage	Primary Metrics	Representative Citations
Screening	Sensitivity (recall), specificity, precision, F1, Cohen's κ	[27–30]
Data extraction	Accuracy, F1-score, Cohen's κ, hallucination rate	[31–34]
Quality appraisal	Cohen's κ, ICC, macro F1, confidence-stratified accuracy	[17,35,36] [a]

[a] Srinivasan et al. evaluate reporting quality (CONSORT compliance) rather than traditional risk-of-bias appraisal.

3. Corpus Characterisation

3.1. Selection Process

The study selection process described in Section 2.5 yielded 122 studies from an initial 1786 records across three databases. The PRISMA 2020 flow diagram is presented as Figure 5.

3.2. Temporal Distribution and Growth

The corpus spans five calendar years: 2022 ($n = 4$), 2023 ($n = 12$), 2024 ($n = 27$), 2025 ($n = 63$), and 2026 ($n = 16$; partial year to March 2026). The year-on-year growth rate between 2022 and 2025 is 200%, 125%, and 133% respectively, confirming the exponential acceleration noted qualitatively in the introduction (Figure 6). Studies published in the first quarter of 2026 alone ($n = 16$) already exceed the entire output of 2023 ($n = 12$), suggesting that the trajectory has not yet plateaued. The sharp post-2024 acceleration corresponds to the public release of GPT-4o (May 2024), Claude 3 (March 2024), and Gemini 1.5 (February 2024), each of which lowered the technical barrier to LLM integration into review workflows. Agentic end-to-end frameworks are disproportionately concentrated in 2025–2026: of the seven studies classified as Stage G (end-to-end agentic), five were published after January 2025, consistent with the hypothesis that agentic capability requires the long-context and tool-use features that became generally available only in 2024.

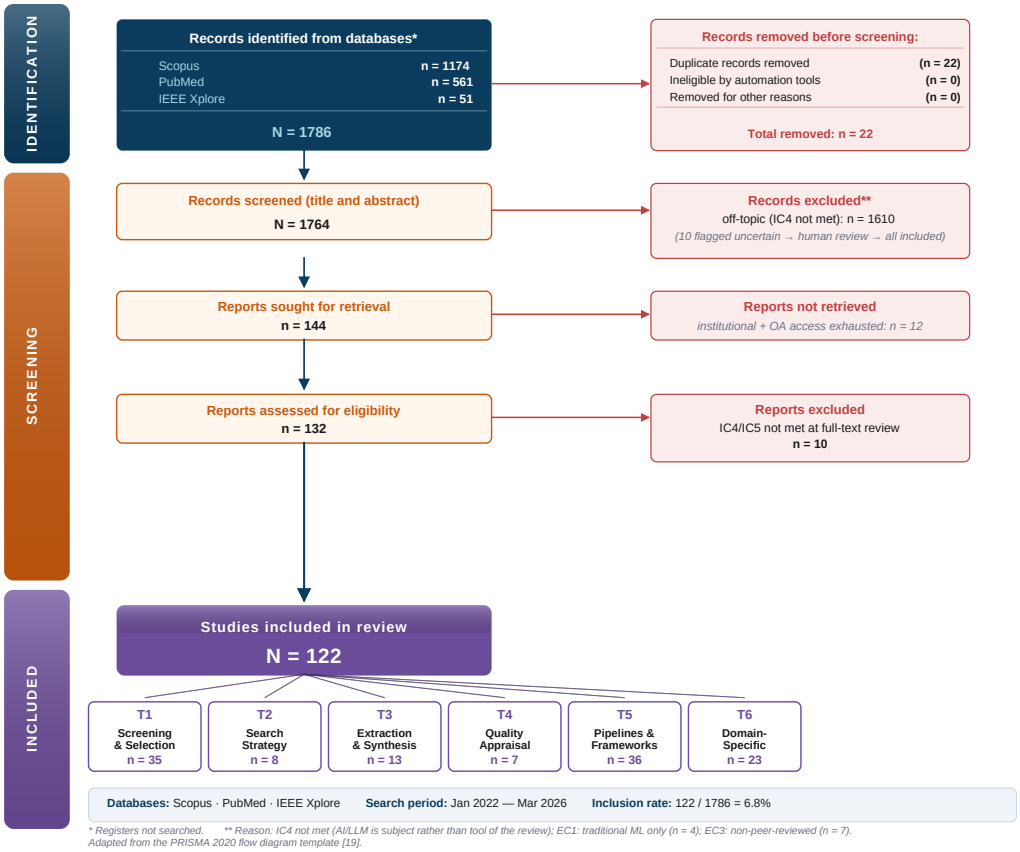


Figure 5. PRISMA 2020 flow diagram for study selection [19]. Records identified from Scopus ($n = 1174$), PubMed ($n = 561$), and IEEE Xplore ($n = 51$); 10 UNCERTAIN records reviewed by human author team (all 10 included); 22 cross-database duplicates removed; 12 records excluded at full-text stage (PDF not retrievable); final corpus $N = 122$.

3.3. Corpus Taxonomy by Review Stage

Table 7 classifies each corpus paper against the seven-stage taxonomy defined in Section 2.1. Multi-stage papers appear in more than one cell. Stage B (Screening) is the most densely populated, containing approximately 50 distinct paper–stage assignments (41% of corpus), followed by Stage E (Extraction, 38 assignments, 31%), Stage A (Search, 28 assignments, 23%), Stage F (Synthesis & reporting, 15 papers, 12%), Stage D (Quality appraisal, 7 papers, 6%), and Stage G (Agentic pipelines, 7 papers, 6%); Stage C (Full-text selection) is reported within screening and not separately enumerated. The asymmetry is consistent with the theoretical argument in Section 1.1: screening is the most tractable SLR stage for LLM automation because it reduces to a classification problem with a binary outcome and a clear gold standard, whereas quality appraisal and synthesis require expert judgement that is harder to formalise as an LLM inference task.

Figure 7 maps the 122 papers across all seven stages and six thematic clusters, making the stage asymmetry immediately visible: Stage B (Screening) is the most densely populated cell across all clusters, while Stage D (Quality Appraisal) remains sparse throughout.

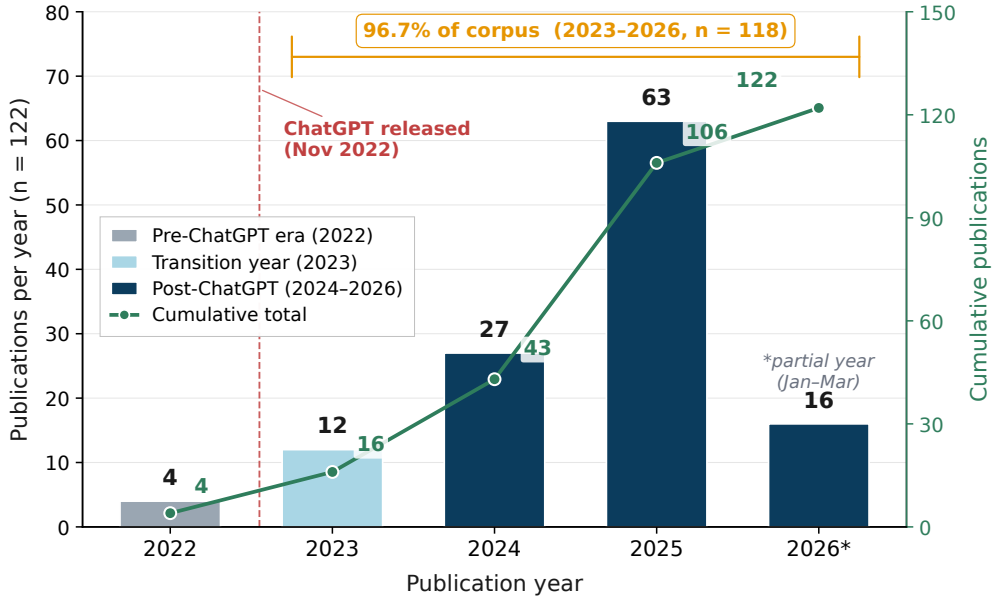


Figure 6. Annual growth of AI-assisted systematic review publications in the corpus ($N = 122$). Bars show publications per year; the dashed line marks the ChatGPT release (November 2022) as the field inflection point; the teal line shows the cumulative total.

Table 7. Corpus taxonomy by review stage. Multi-stage papers appear in more than one row. Stage codes (canonical, per Figure 1): A = Search; B = Screening; C = Full-text Selection; D = Quality Appraisal/Risk of Bias; E = Extraction; F = Synthesis & Reporting; G = Agentic Pipelines (≥3 stages).

Stage	Representative Papers (Selected)
A—Search/query generation	[12,37–50]
B—Screening/study selection	[1,5,6,13–15,17,18,26–30,48,51–85]
C—Full-text selection	Reported within screening (Stage B); not separately enumerated.
D—Quality appraisal/risk of bias	[35,36,61,86–92]
E—Extraction/charting	[3,6,13,15,31,50,53,56,89,93–100]
F—Synthesis & reporting	[2,24,49,96,101–105]
G—Agentic pipelines (≥3 stages)	[2,8,86,90,106]

3.4. LLM Families and Evaluation Designs

Table 8 distributes the corpus by LLM family. OpenAI's GPT series dominates (49 of 122 studies), consistent with a contemporaneous scoping review finding GPT-family models in 89% of LLM-assisted review evaluations [9]. Open-source models (LLaMA 2/3, Mistral, Vicuna, Falcon, BERT variants, Galactica) appear in 16 studies; their impetus is reproducibility and cost, deployed locally [69] or via free-tier APIs [62]. Claude (Claude 2–3.7 Sonnet) appears in 10 studies and is consistently strong on structured extraction: in the six-model benchmark of Sollini et al. [10], Claude 3.7 Sonnet was the only model to produce a proper reference section during drafting; Silva and Wickramaarachchi [107] similarly highlight Claude.ai and ChatGPT as structured-summary producers across ten evaluated tools. Gemini appears in six studies [15]; early Gemini/Bard evaluations were unfavourable (91.4% reference-list hallucination rate vs. 28.6% for GPT-4 and 39.6% for GPT-3.5 [108]), but recent versions have improved—Gemini 2.5 Pro with Deep Research identified 13 of 18 eligible papers with zero hallucinations in the same benchmark. DeepSeek appears in three (all post-2024) studies, principally for text-understanding and extraction [50].

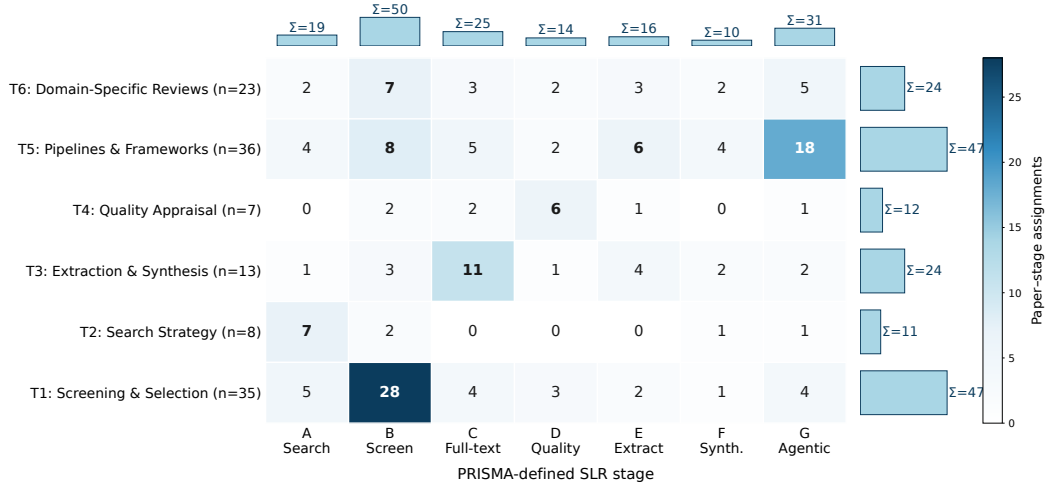


Figure 7. Distribution of the 122 included studies across PRISMA-defined review stages (columns) and thematic notebook clusters (rows). Cell values indicate paper–stage assignments; column and row margins show stage and cluster totals. The dominance of Stage B (screening) and the near-absence of Stage D (quality appraisal) in most clusters reflect the field's structural automation asymmetry.

Evaluation designs split as follows: case-study/single-review retrospective designs dominate (68 studies, 56%); multi-review benchmark evaluations account for 29 studies (24%), including Delgado-Chaves et al. [18], Syriani et al. [71], and the contamination-aware framework of Kusa et al. [109]; prospective and near-prospective designs account for 16 studies (13%), with Gartlehner et al. [93] advocating the Study Within A Review (SWAR) design, Scherbakov et al. [8] deploying a custom LLM plugin in an ongoing review, Homiar et al. [27] evaluating workload reductions retrospectively against living systematic review updates, and Sung et al. [70] retrospectively evaluating ranking reliability and review workload against recently published Cochrane systematic reviews. The remaining nine (7%) are tool demonstrations without a quantitative comparator.

Table 8. LLM family and evaluation design distribution across the corpus ($n = 122$). Multi-model papers appear in more than one LLM family row. Percentages computed over the 122 corpus papers.

LLM Family	*n*	Selected Papers
GPT (OpenAI)—GPT-3.5, GPT-4, GPT-4o, GPT-4o-mini	49	[1–3,6,8,10,11,15,17,18,24,26–28,30,37,40–45,48–51,54,55,57,58,61–67,70,77,78,86,88,91,94,98,106,110–112]
Open-source (LLaMA, Mistral, Vicuna, Galactica, BERT variants)	16	[2,10,13,18,24,37,55,59,61,62,69,73,75,91,98,112]
Claude (Anthropic)—Claude 2, Claude 3, Claude 3.5/3.7 Sonnet	10	[9,10,26,28,30,43,51,55,77,107]
Gemini (Google)—Gemini 1.5, 2.0, 2.5	6	[5,10,15,30,51,108]
DeepSeek	3	[10,50,63]
Proprietary/tool-wrapped (Elicit, Copilot, Consensus, Rayyan, SciSpace)	11	[1,6,8,50,65,74,95,96,107,111,113]
Not disclosed/undisclosed	5	[studies in which the underlying LLM was not identified by authors]

Note. Multi-model studies appear in multiple rows; the total exceeds 122.

3.5. Domain Coverage and Review Type

The corpus spans a broad range of academic disciplines, confirming that LLM-assisted reviewing is not restricted to any single field. Biomedical and clinical science accounts for the largest share (41 papers (34%)), comprising evaluations of LLM pipelines applied to clinical medicine broadly [16,86], methodological tutorials and evaluations covering cardiology [90], orthodontics [114], and LLM-assisted evidence synthesis for burn wound assessment [101]. The concentration in biomedical settings reflects the field's pre-existing

infrastructure for evidence-based practice and the availability of large, publicly accessible screening benchmark datasets used in technology-assisted review evaluation.

Figure 8 illustrates that no single model family achieves uniformly high performance across all review stages, reinforcing the case for hybrid architectures that deploy different models at different pipeline stages rather than a single end-to-end LLM; general-purpose models excel at search and screening, while domain-fine-tuned models lead at extraction and appraisal. Because the underlying studies report heterogeneous metrics (accuracy, recall, F1, Cohen's κ) against different datasets and gold standards, the axes are not pooled onto a common numerical scale; each indicates a model family's qualitative relative strength within a task rather than a directly comparable magnitude (cf. the benchmark-incomparability limitation, Section 9.2). Computer science and information systems is the second largest domain (28 papers (23%)), comprising LLM-assisted screening evaluations [14] and studies treating systematic reviewing as a computational problem. Social sciences, education, and public health collectively account for 22 papers (18%), including evaluations of LLM assistance in second-language acquisition reviews [53], relationship education [15], social work practice [52], qualitative research methodology [97], and disability studies [13]. Engineering and environmental science—including reviews of photovoltaic power plant site selection [50], contaminated soil remediation [115], and vehicle automation energy consumption [116]—constitute the remainder alongside health economics [73] and legal/patent analysis [117].

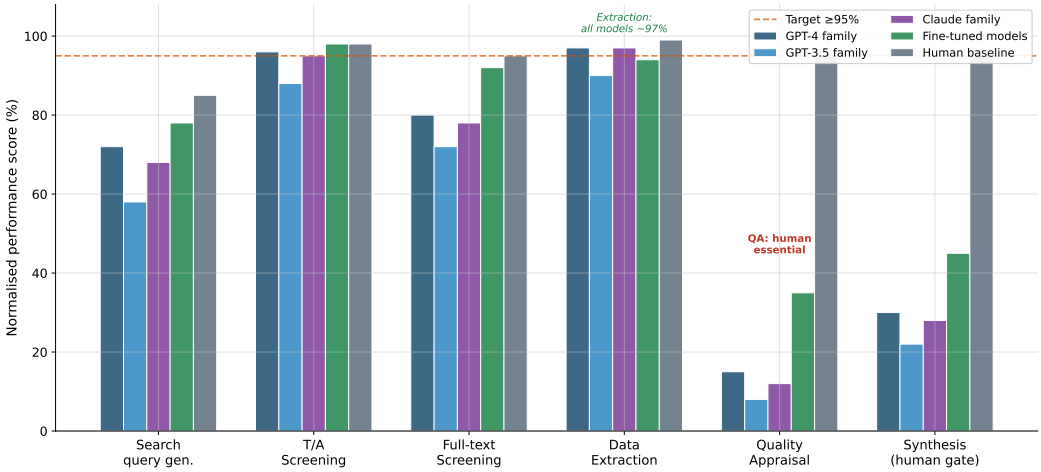


Figure 8. Qualitative, relative performance comparison of large language model families across systematic review task categories: search query generation, title/abstract screening, full-text screening, data extraction, quality appraisal, and synthesis. Axes show relative ordinal positioning, not directly comparable magnitudes (see text and Section 9.2).

Several papers in the corpus are methodological studies, scoping-review protocols, or horizon-scanning analyses that discuss AI adoption in reviews without implementing an AI tool to produce novel empirical findings. Teperikidis et al. [90] provide a comprehensive tutorial on performing systematic literature reviews using ChatGPT 4.0 for cardiologists, presenting a methodological roadmap and theoretical examples across all review stages rather than reporting new clinical findings. Demir et al. [114], however, diverge from these non-empirical reports by conducting a quantitative evaluation of ChatGPT models, specifically measuring their accuracy in automating the initial search stage through PICO formulation and Boolean query generation. Additionally, a scoping review protocol by Ndukwe and Otukpa [118] details a planned methodology for examining how large language models are audited and tailored for use within healthcare education and public health learning spaces. A further variant is represented by papers that are themselves

systematic reviews of AI applications in a specific domain, conducted using traditional human-led methods: Jung and Jin [119] synthesise 28 studies on the role of generative AI as a collaborator in educational knowledge construction, characterising how students and educators leverage generative AI tools in learning environments, without deploying AI for any stage of the review process itself.

Of the 116 empirical papers, 74 (64%) evaluate a single SLR stage in isolation—including studies where a planned second-stage AI evaluation was abandoned due to poor performance [15]; 20 (17%) evaluate two stages using AI-assisted approaches, whether as separate independent tasks or integrated pipelines [53,56]; and 22 (19%) evaluate three or more stages, though these remain human-directed augmented pipelines rather than fully autonomous end-to-end agentic systems. This distribution underscores the principal structural gap in the literature: most empirical work addresses the SLR pipeline as a series of disconnected sub-problems rather than as an integrated system.

4. Search-Query Generation and Screening

4.1. LLM-Generated Search Strategies

The eight papers in the search-strategy sub-corpus (T2) collectively address a foundational question: can a large language model construct or refine a systematic review search string to the standard required for PRISMA-compliant retrieval? The empirical answer is markedly sensitive to the evaluation design, the database targeted, the prompting strategy deployed, and crucially, the comparison baseline—with every study finding that LLM-generated queries underperform librarian-authored strategies on recall, even as they offer substantial savings in expert labour.

Evaluating a LoRA-fine-tuned Mistral-Instruct-7b model, Adam et al. [37] demonstrate that it achieves an 86% median sensitivity against a librarian-authored gold standard, though at a much lower precision, producing a number needed to read (NNR) of 908 compared to an NNR of 580 for human experts. Systematic prompt-strategy comparisons reveal a trajectory of significant improvement: while initial evaluations found that the best-performing prompts achieved an average recall of only 0.50 to 0.52 [42], a recent reproducibility study across ten post-2023 models demonstrates that advanced reasoning architectures (such as OpenAI-o1) have significantly narrowed the gap to approximately 15% behind manual human queries [43]. Knowledge-base augmentation [41] improves median recall to 69.2% (IQR: 43.8–88.9%) across two databases (PubMed and Embase), while PRESS-framework evaluations [44] show that GPT-4 and GPT-3.5 scored identically in an initial round of zero-shot testing, but GPT-4 outperformed GPT-3.5 on syntax and controlled vocabulary in a second round of iterative evaluation, with both models still requiring human expert review. A critical practical limitation is session-to-session irreproducibility: Park et al. [40] document unstable keyword inventories across independent sessions given identical tasks. Per-study metrics are provided in the Supplementary Materials (Table S5).

Failure Modes in LLM-Generated Search Strategies

Five recurrent failure modes are documented across the T2 sub-corpus: hallucinated or misapplied controlled vocabulary (e.g., non-existent MeSH terms) [38], database-syntax violations (unclosed brackets, invalid Boolean sequences), inappropriate over-narrowing via unrequested date/language/design filters, omission of vital synonyms, and overmatching through indiscriminate subject-heading explosion. The convergent implication is that no current configuration produces a search string suitable for independent use without expert information-specialist review: post-generation peer review and guided seed-study formulation are necessary safeguards rather than optional checks.

4.2. Title-and-Abstract and Full-Text Screening

The screening sub-corpus (T1, 35 papers) constitutes the most densely evidenced stage in the corpus, accounting for approximately 41% of the corpus (≈50 paper–stage assignments; Section 3.3). This concentration of research effort reflects the structural tractability of screening as an LLM inference task: it reduces to a binary classification problem (include/exclude) with a well-defined gold standard in the form of the final included study set.

4.2.1. Headline Performance and Determinants

LLMs reach 100% sensitivity in a variety of configurations—conservative inclusion instructions [58], low-rating thresholds, ensemble majority voting, and zero-shot prompting [84]—though peak sensitivity is sometimes inflated by benchmarking against post-full-text inclusions rather than the initial corpus [27]. The cost is precision: it regularly falls below 20% because of false-positive volume (Supplementary Table S1). Agreement metrics span a wide continuum: Cohen's κ ranges from 0.27 on average [29] and 0.35–0.43 [63] through 0.79 at title/abstract and 0.61 at full-text screening for a fine-tuned model [78], up to 0.69 in broader searches [82]; F1 ranges from 0.27 (abstract) to 0.90 (full-text) under short-prompt conditions [28]. Specificity is sensitive to model generation—GPT-4 outperforms GPT-3.5 by up to 47 pp on the same corpus [30,64]—to prompt design [66,71], and to domain and pretraining [68]; NPVs, in contrast, consistently exceed 99.5% and frequently reach a perfect 100% under optimal prompting conditions [28,58], providing the structural basis for the deployment consensus in Section 4.3. When LLMs assist rather than replace reviewers (PICOS-extraction for human adjudication), human–machine collaboration reaches 100% sensitivity, 99.9% specificity, and $\kappa = 0.98$ (PABAK 0.998), allowing junior researchers to outperform unassisted experts [72]. The challenge is therefore not whether LLMs can screen, but under what conditions and configurations they do so safely.

4.2.2. Meta-Analyses, Ensembles, and Large-Scale Evaluations

The one formal meta-analysis [17] pools 14 LLMs and finds sensitivity and specificity converging near 90%, with near-perfect sensitivity requiring a steep specificity trade-off. Delgado-Chaves et al. [18] compared 18 models across three reviews and identified GPT-4o, GPT-4o-mini, and GPT-3.5-turbo as the most balanced configurations, reducing human workload by up to 93% on the recommendation that LLM exclusions be trusted only when confirmed by an independent human reviewer. Wang et al. [75] show that a calibrated ensemble of zero-shot open-source models (LLaMA, Falcon, Alpaca variants) reliably meets pre-set recall targets without task-specific fine-tuning. Ensembling more generally maximises recall and offsets individual model biases [51]: a three-model majority vote (GPT-4o-mini + Llama 3.1 8B + Gemma 2 9B) reaches 93.1% balanced accuracy at a 50-rating threshold [69], and a post hoc Claude/Gemini union ensemble achieves 99.0% sensitivity with 70.0% specificity [67], with union strategies systematically trading specificity for recall.

4.2.3. Calibration and Real-World Performance

LLM screening performance is not fixed at a single operating point but highly sensitive to scoring thresholds. Seed-study calibration uses token probabilities to pre-specify and guarantee target recall levels [75]; a complementary one-shot "inclusivity adjustment" counts inclusion/exclusion criteria per abstract without re-inferring the full corpus [61]; and Likert/0–100 rating scales enable granular threshold manipulation [29,69]. Retrospective simulations document the brittleness of these operating points under real-world conditions: LLM sensitivity drops to 58% between update rounds of a living review [27], and replicating 25 Cochrane reviews shows that some included articles are ranked so poorly that up to

96% of the corpus would require manual screening to recover them [70]—confirming that human review of all abstracts remains necessary. A counterpoint is Scherbakov et al. [8], whose custom LLM plugin for a methodological review substantially reduced manual workload without loss of accuracy.

Figure 9 plots the sensitivity–specificity operating points for all documented configurations. Fine-tuned models cluster in the target operating zone, whereas general-purpose models achieve high sensitivity at the cost of low specificity; the pronounced horizontal spread at high sensitivity confirms that calibration—not raw model capability—is the primary determinant of practical utility.

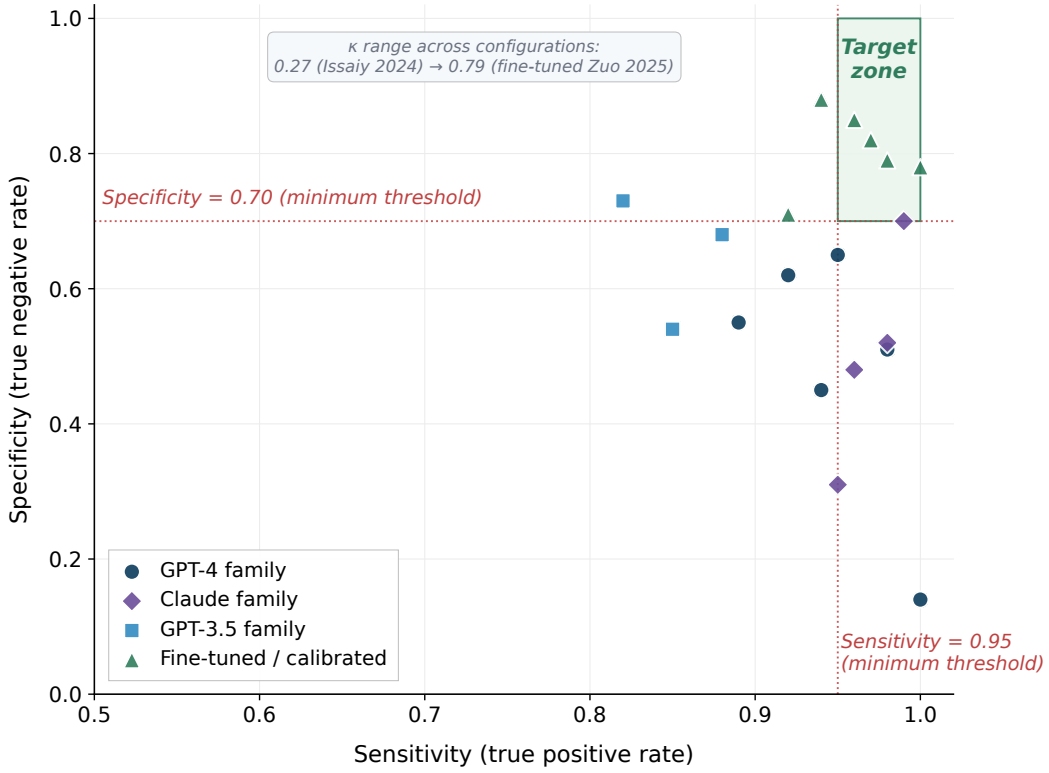


Figure 9. Sensitivity–specificity trade-off across LLM-assisted title/abstract screening configurations ($n > 20$ configurations from the T1 sub-corpus). Point colour indicates model family; the shaded rectangle marks the target operating zone (sensitivity ≥ 0.95, specificity ≥ 0.70).

4.3. Safety, Deployment Conditions, and Systematic Failure Modes

4.3.1. The Deployment Consensus and Failure Modes

The consensus across the screening literature is unambiguous: LLMs are not yet safe for fully autonomous, unsupervised deployment and must serve as triage tools, first-pass filters, or second reviewers in a hybrid human–AI workflow. The conclusion rests on two structural findings: the near-universal failure of LLMs to match dual-reviewer consensus reliability at title/abstract [77,82] and systematic over-inclusion that inflates manual workload (Section 4.2). The consensus does not forbid trusting LLM exclusions: at NPVs of 99–100%, efficiency is achieved precisely by trusting exclusions while concentrating human verification on the high-false-positive include pool [58] and on disagreements between the LLM and a primary human reviewer.

Three systematic failure modes recur. *Information-extraction deficits*: LLMs fail when eligibility-relevant data are not stated explicitly in the abstract, tending to over-exclude because they conflate abstract with full record [48,61,64]. *Ambiguity paralysis*: disagreements among multiple LLMs and between LLMs and human reviewers cluster around borderline abstracts with insufficient context or domain-specific terminology [27,68]—useful as a

routing signal for hybrid pipelines. *Over-inclusivity under sensitivity-maximising prompts*: sensitivity-optimised prompts [30], criteria omissions [69], or union-ensemble aggregation [67] flood the full-text stage with false positives—addressable by threshold calibration but ultimately a precision–recall trade-off unresolved by prompt engineering alone.

4.3.2. Agentic Screening Pipelines

A subset of T1 papers describe screening within multi-stage pipelines that incorporate automated handoff between stages. Nordmann et al. [65] evaluated ChatGPT 4.0's feasibility for actual title and abstract screening, finding that despite achieving 88–89% sensitivity and substantial workload savings, the model produced an 11% false-negative rate and demonstrated only "slight" inter-rater agreement ($\kappa = 0.12$–0.13) with human reviewers, making human oversight paramount. Furthermore, Nordmann and Fischer [66] demonstrated that while ChatGPT possesses "substantial" internal consistency across repeated screening rounds ($\kappa = 0.76$), its reasoning about inclusion rules remains highly sensitive to how humans structure the criteria, requiring researchers to carefully combine differently scoped criteria to safely identify ambiguous cases for human review. These findings establish that even the preparatory stage preceding screening itself—eligibility-criteria formalisation—requires human oversight, reinforcing the conclusion that fully autonomous agentic screening pipelines remain a medium-term rather than near-term prospect.

Offering an alternative design that avoids black-box AI classification entirely, Robleto and Shehadeh [120] deploy Claude 4 and ChatGPT strictly as coding assistants to write and modify Python 3.12.7 keyword-matching scripts, which then execute rule-based two-phase screening—title/abstract followed by full text—without direct LLM inference on abstracts; this workflow eliminates machine-learning dependencies, reduces API costs, and produces a fully auditable screening trail, which was demonstrated across cardiac fatty acid oxidation and e-cigarette cardiopulmonary case studies. Giunti and Doherty [121] adopt a comparable coding-assistant model, using conversational exchanges with ChatGPT to co-create a Python script that automates keyword-searching and data extraction for a systematic review of mobile health applications from the Google Play Store—demonstrating that LLMs can serve as collaborative programming tools to build custom corpus-assembly pipelines for researchers with limited coding skills, without delegating any classification decisions to the model.

5. Data Extraction and Quality Appraisal

5.1. Structured Data Extraction from Primary Studies

Data extraction—the stage at which reviewers abstract standardised fields from full-text sources—has attracted the second largest body of work in the corpus, with approximately 31% of paper–stage assignments (Section 3.3). The 13 papers in the T3 extraction sub-corpus span PICO field extraction, numerical outcome abstraction, clinical data charting, and machine-translated record processing, collectively permitting a nuanced characterisation of where LLM-based extraction reliably succeeds and where it characteristically fails.

5.1.1. Headline Accuracy and Field-Level Variation

The most frequently cited extraction benchmark is that of Gartlehner et al. [31], who used Claude 2 with zero-shot, plain-language queries to extract 16 data elements from 10 randomised controlled trials included in a systematic review of treatments for plaque psoriasis. Prompts were iteratively engineered to specify exact output constraints (e.g., requesting numerical data reported to specific decimal places). Overall accuracy was 96.3% (F1 = 0.98), with a human-extracted reference standard verified by a second investigator against the source PDF. However, accuracy varied markedly by field type: 100% for

basic study metadata (first author, study name, funder) and for participant flow counts, but dropping to 80% for primary outcome effect estimates. The study recorded six errors in total, of which two (33.3%) were hallucinations—one fabricating an exclusion criterion and one fabricating a primary outcome estimate—while a companion study (Konet et al. [122]) testing GPT-4 on the same 10 trials produced one hallucination out of 49 total errors—though 35 of these errors (71.4%) were explicitly caused by the third-party PDF-parsing plug-in required for document access rather than the language model itself [122].

Motzfeldt Jensen et al. [32] evaluate ChatGPT-4o on a systematic review of exercise interventions to reduce fall risk. The model was highly accurate for Intervention and Comparator fields (99.2%), General Information (98.2%), and Population (92.9%), but accuracy dropped markedly to 73.9% for Outcome data. The hallucination rate across all fields was 5.2% (95% CI: 3.4–7.4%), meaning that approximately 1 in 20 extracted data points contained a value entirely absent from the source document. The model most commonly hallucinated when specific data points were not explicitly reported in the article, prompting it to fabricate plausible-looking values rather than flag a missing-data entry.

Figure 10 presents both the field-level extraction accuracy profile (Panel A) and the quality appraisal agreement landscape (Panel B), confirming that while extraction is broadly reliable for most field types, quality appraisal agreement remains near-chance across all evaluated instruments.

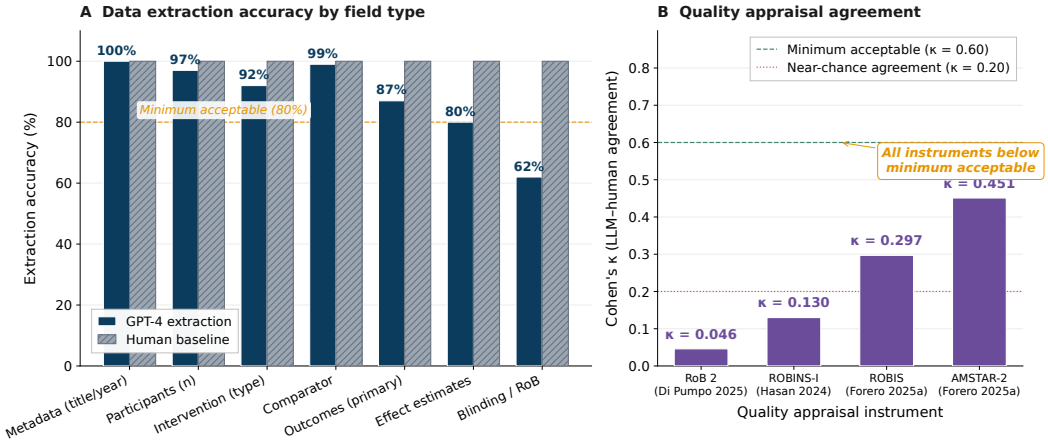


Figure 10. LLM performance in data extraction and quality appraisal. (**A**) Field-level extraction accuracy (%) for GPT-4 and Claude relative to the human baseline (dashed line); effect estimates are the most error-prone field. (**B**) Cohen's κ for four quality appraisal instruments; the red dashed line marks the minimum acceptable threshold ($\kappa = 0.60$).

Khan et al. [123] conducted a multi-model comparison of GPT-4-turbo and Claude-3-Opus on extraction tasks with cross-critique of discordant responses. Hallucinations were present in 2.3% of GPT-4-turbo responses and 2.8% of Claude-3-Opus responses. Analysis of hallucination mechanisms revealed a qualitative distinction between the two models: GPT-4-turbo hallucinated primarily due to misinterpretation at the variable level (62% of its hallucinations), whereas Claude-3-Opus hallucinated primarily due to extracting from the wrong section of the document (56% of its hallucinations), suggesting that because these models possess distinct processing patterns, they offer complementary perspectives that make a collaborative, cross-critique workflow highly effective at mitigating errors overall.

5.1.2. Schema-Constrained Decoding and Prompt Engineering

Several extraction studies demonstrate that constraining the model's output format reduces error rates substantially relative to free-text generation. Addressing the context-window barrier that constrains full-text extraction, Landschaft et al. [124] implement a

GPT-4 pipeline with retrieval-augmented generation (RAG) via LangChain—splitting full-text PDFs into 1200-character chunks stored in a ChromaDB vector database and retrieved via multi-query prompts—achieving near-perfect inter-rater agreement at the abstract-screening stage (Cohen's $\kappa = 0.890$–0.971; F1 = 0.950) but observing a performance drop during full-text extraction (accuracy = 0.870; F1 = 0.769). Per-study accuracy results for the 13 T3 extraction studies are provided in the Supplementary Materials (Table S6); Section 5.2 details the field-level failure modes that structured output constraints most directly address.

5.2. Extraction Failure Modes: Complex Tables, Supplementary Materials, and Qualified Claims

Three failure categories recur across the sub-corpus. *Complex tables*: models omit available information (trial registry numbers, inclusion criteria) or occasionally hallucinate primary outcome estimates [31], and full-text segmentation to bypass character limits disrupts textual coherence [94]. *Supplementary Materials and companion publications*: when tools cannot access appendices, information is incorrectly marked as "not available" [125]; TrialMind produced 10 such "unavailable data" errors alongside only three hallucinations, with the authors framing this positively as evidence of robustness [43]. *Conditional, qualified, or implicit claims*: Kasireddy et al. [33] document a 10% error rate on "Estimate Value" (arm-level versus population-level data) and a further 10% on "Blinding" (defaulting to "open label" rather than "not reported"); Liu et al. [126] report 62.24% of participant-number errors arise from failure to infer implicit data. The primary treatment of hallucination rates and mitigations is Section 6.2.

5.3. Risk-of-Bias and Quality Appraisal

Quality appraisal—the formal assessment of methodological risk of bias using validated instruments—has received the least automation effort of any PRISMA-defined stage: only approximately 6% of papers in this corpus address quality appraisal as a primary focus (Section 3.3). Of the seven papers with a Stage D assignment, five are concentrated in the T4 thematic notebook because their primary contribution is quality-appraisal tool evaluation; the remaining two address quality appraisal as part of broader multi-stage pipelines and are analysed in Section 7. The five papers in the T4 quality-appraisal sub-corpus evaluate LLMs against the Cochrane Risk of Bias 2 (RoB 2) tool, the ROBINS-I instrument for non-randomised studies, AMSTAR-2 for systematic reviews of systematic reviews, GRADE evidence profiling, and domain-specific quality checklists, uniformly reporting performance substantially below the levels documented for screening and extraction.

5.3.1. Agreement with Human Assessors and Instrument-Specific Performance

Table 9 summarises the five corpus studies that evaluated LLMs for risk-of-bias and quality appraisal. Di Pumpo et al. [87] provide the most granular empirical analysis, evaluating ChatGPT-4 deployed as custom GPTs against a reference review's original RoB 2 and Levels of Engagement Reached (LOER) classifications for 36 RCTs with a double-repetition verification strategy. RoB 2 agreement was 47.2% (17/36, $\kappa = 0.0455$; per-domain $\kappa \in [-0.19, 0.21]$); LOER agreement was 27.8% ($\kappa = 0.0445$). A systematic central-tendency bias accompanies these low κ values: for RoB 2, the model classified only 13.9% of studies as "high risk" vs. 25.0% by humans, overrepresenting "some concern"; for LOER, the bias reversed, with 91.7% placed in the "Collaborate" tier vs. 25.0% by humans. Hasan et al. [88] report a consistent pattern on ROBINS-I: overall $\kappa \approx 0.13$, with the highest raw agreement on intervention classification (71%) and the lowest on deviations from intended interventions (31%) and selective reporting (36%). Per-domain κ values ranging 0.02–0.28 confirm that LLMs cannot currently replace independent human reviewers for complex risk-of-bias assessments.

Table 9. LLM studies for risk-of-bias and quality appraisal (n = 5). RoB 2 = Cochrane Risk of Bias 2; ROBINS-I = Risk Of Bias In Non-randomised Studies; AMSTAR = A MeaSurement Tool to Assess systematic Reviews; LOER = Levels of Engagement Reached; CT = central tendency.

Author (Year)	LLM	Instrument	Agreement (κ)	Calibration
[87]	ChatGPT-4 (custom GPT)	RoB 2; LOER	RoB 2: $\kappa = 0.0455$; LOER: $\kappa = 0.0445$	CT bias: under-risk, over-engagement
[88]	GPT-4	ROBINS-I	Low overall ($\kappa \approx 0.13$); highest on intervention classification (71%); lowest on deviations from intended interventions (31%)	Low κ likely reflects skewed distribution of ratings
[36]	GPT-4o-mini	CONSORT 2010 (human reporting compliance)	F1 = 0.86; precision = 0.97 as evaluator of human reporting quality	High precision with lower recall; evaluates human authors, not LLM appraisal
[35]	QWEN; Gemini 2.0 Flash; ChatGPT-4o; DeepSeek	ROBIS; AMSTAR 2	ROBIS: $\kappa = 0.297$ (Gemini, best); AMSTAR-2: mean $\kappa = 0.451$ across 4 LLMs (range 0.169–0.605)	Central-tendency bias (QWEN defaults to "moderate" for 90% of ROBIS answers); AMSTAR-2 varies widely by model
[90]	ChatGPT (version unspecified)	AMSTAR 2	Tutorial coverage; quantitative evaluation not reported	Guidance provided; calibration not formally measured
[92]	GPT-4	AMSTAR 2	High internal consistency (85%); κ not reported	Hallucination rate "moderately low"

5.3.2. Calibration Behaviour and Confabulated Rationales

The central-tendency bias above is confirmed independently: Forero et al. [35] report QWEN defaulting to "moderate" for 90% of its ROBIS answers regardless of actual study quality, paralleling the LOER and RoB 2 patterns of Di Pumpo et al. [87]. On fabricated rationales—LLMs producing justifications not supported by the study text—Yao et al. [92] characterise hallucination in quality appraisal as a "well-described" but "moderately low" risk, with 85% internal consistency across iterations.

6. Evidence Synthesis, Reporting, and Transparency

6.1. Narrative Synthesis and Meta-Analysis Preparation

Evidence synthesis—aggregating, interpreting, and writing up extracted evidence across included studies—represents approximately 6% of paper–stage assignments in the corpus (Section 3.3) and is the stage where the gap between LLM capability and the expert reviewer standard is most consequential: a hallucinated synthesis claim propagates through the entire review product and may influence clinical or policy decisions. The corpus ranges from highly targeted synthesis deployments to full manuscript automation, with the breadth of applications already spanning every narrative writing task in a systematic review.

Applications range from conservative, bounded uses to near-complete pipeline automation. At the conservative end, Lieberum et al. [9] deployed Claude 3.5 Sonnet exclusively for writing plain-language summary sections, while Holm et al. [101] used an LLM to draft a literature-synthesis orientation decision tree subsequently reviewed by two consultant burn surgeons. These targeted applications represent the deployment model most synthesis-stage reviewers follow: LLMs are inserted into a single, well-bounded subtask with human experts retaining decision authority.

At the opposite end, pipeline systems generate structured LaTeX review documents or draft complete manuscript sections across multi-model benchmarks; even in these high-automation designs, human review before publication remained mandatory (Section 7).

The most practically important synthesis contribution concerns meta-analytic preparation. Purewal et al. [11] documented near-perfect accuracy in calculating pooled effect estimates and observed no gross discrepancies upon visual inspection of the generated forest plots—in sharp contrast to the same study's 54.9% abstract-screening sensitivity (Section 4.2), illustrating the stage-specificity of LLM capability.

Among early deployments, Alshami et al. [106] established a template for automated content analysis, using ChatGPT to systematically extract and categorize trends, limitations, and research gaps, while human editors retained responsibility for evaluating the outputs and synthesizing the final narrative. Qualitative synthesis remains more challenging: although LLMs can generate concise summaries, they struggle to cross-reference context-dependent and interpretive information scattered throughout qualitative manuscripts [96], and they generally lack the critical interpretative reasoning of domain experts required for nuanced synthesis [2,10], a limitation consistent with the epistemological critique of Erslev and Tretow-Fish [6]. One mitigation for such limitations is pairing LLMs with domain-appropriate statistical tools: da Costa Mineiro et al. [127] integrate ChatGPT with R's Bibliometrix package and ChatGPT to construct a structured analysis roadmap for a bibliometric systematic review of 1545 innovation-ecosystem articles, with ChatGPT providing accessible descriptive summaries and initial theoretical orientation while R handles the quantitative analysis and visualisation that LLMs lack—demonstrating that AI and statistical software serve complementary rather than redundant functions in the synthesis stage.

6.2. Hallucination, Grounding, and Human Post-Editing

The hallucination risk in synthesis prose is qualitatively different from that in structured extraction: a fabricated extraction value is detectable by cross-referencing the source document, but a fabricated synthesis claim that correctly cites a real paper while mischaracterising its conclusion may be undetectable without re-reading every source. The T3 subcorpus provides the most precise estimates currently available. Motzfeldt Jensen et al. [32] documented a 5.2% overall hallucination rate (95% CI: 3.4%–7.4%) on 484 data points from 11 RCTs (ChatGPT-4o), with hallucinations spiking when requested data were absent from the source. Khan et al. [123] reported 2.3% for GPT-4-turbo (62% variable-level misinterpretation) and 2.8% for Claude-3-Opus (56% context-level misinterpretation) across 23 variables in 22 publications.

Figure 11 makes the cross-stage disparity apparent: extraction hallucination at 0.6–5.2% is containable, whereas synthesis-stage citation hallucination is an order of magnitude higher and unmitigable without a mandatory human post-editing gate. Earlier benchmarks agree: Gartlehner et al. [31] evaluated Claude 2 on 160 data elements from ten RCTs (six errors, two hallucinations); Konet et al. [122] directly compared Claude 2 and GPT-4 on the same dataset (GPT-4: 49 errors, finding one hallucination among 160 data elements, representing a 0.6% hallucination rate). These rates propagate directly into pooled effect estimates, and the corpus consistently concludes that a human post-editing gate is essential [100,122,125]; isolated testing of ChatGPT on medication-efficacy synthesis produced fluent but factually ungrounded summaries [102].

Two complementary mitigations reduce error rates measurably. Schema-constrained decoding: Kasireddy et al. [33] minimise false positives in a clinical extraction pipeline via Claude 3.7 Sonnet with strict JSON-schema validation, role-based prompts, a domain master list, and zero-temperature deterministic output; Kaur and Singh [99] use zero-shot JSON-format constraints with LLaMA 3.3-70B to normalise PICO elements into canonical forms. Human-in-the-loop arbitration: Hilkenmeier et al. [125]'s hybrid dual-extraction architecture flags only discordant data points for reconciliation, reaching 100% accuracy

where AI and human agree and bypassing manual secondary review for ∼52% of data points; García-Torres et al. [128] deploys multiple models in parallel and reconciles discordant outputs through a human adjudicator. No corpus study reports a synthesis or extraction pipeline achieving acceptable accuracy without any human verification step: the post-editing gate cannot be removed from any synthesis pipeline intended for peer-reviewed publication until automatic hallucination detection matches human expert review.

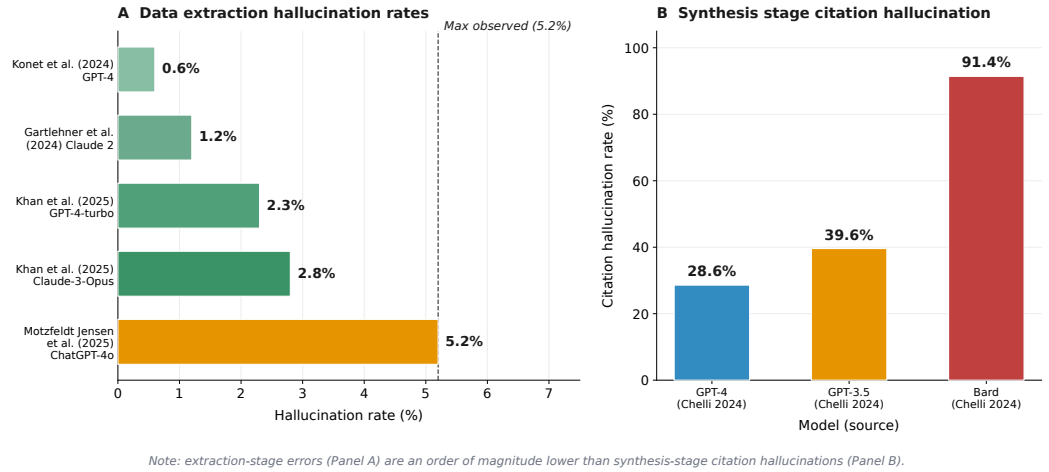


Figure 11. Hallucination and error rates across LLM-assisted systematic review stages. Panel (**A**): data extraction hallucination rates range from 0.6% to 5.2% across five LLM configurations [31,32,122,123]. Panel (**B**): citation hallucination rates at the synthesis stage are an order of magnitude higher (GPT-3.5: 39.6%; GPT-4: 28.6%; Bard: 91.4%) [108].

6.3. Reporting, Transparency, and Reproducibility

The reporting stage—writing and formatting the final review manuscript—has attracted the least automation effort of any PRISMA-defined stage: approximately 7% of corpus papers address reporting as a primary focus (Section 3.3). The T5 sub-corpus additionally reveals that AI transparency and reproducibility practices vary enormously across the corpus, from complete non-disclosure to the most detailed methodology reporting yet observed in the field.

6.3.1. PRISMA 2020 Compliance and Disclosure Practices

PRISMA 2020 [19] contains no items specifically mandating AI-assisted-methodology disclosure. In the T6 sub-corpus, out of 22 primary reviews, only Chen et al. [86] explicitly claims PRISMA 2020 compliance (with acknowledged deviations: no prospective registration, no protocol, no formal RoB), 16 achieve partial compliance, and 2 registered prospectively. Teperikidis et al. [90] provide an end-to-end ChatGPT tutorial covering search through write-up; Soprano et al. [25] map their AI-aided scoping review to PRISMA 2020 items 4–10b, 13a, 16a, 20a–d, 21, 22.

Prompt disclosure is consistently partial or absent across the T5 sub-corpus. Masinde [49] is exceptional in disclosing exact prompts in a dedicated "Declaration" section, and Alshami et al. [106] provides verbatim screenshots of ChatGPT responses; Lieberum et al. [9] gives a formal transparency declaration but no verbatim prompts. Model-version reporting ranges from precise ("ChatGPT-4, subscription-based version available in February 2024" in Di Pumpo et al. [87]) to generic identifiers providing no reproducibility information. Temperature and top-p are disclosed in only two corpus papers: Tripathi et al. [5] (temperature = 0.1, top-p = 0.95) and Yun et al. [24] (top-p = 0.7, max_length = 2048).

6.3.2. Emerging Reporting Frameworks and the Transparency Gap

The transparency data in Table 10 define a reproducibility crisis in AI-assisted systematic reviewing: a review using an unspecified LLM with undisclosed prompts at unknown temperature cannot be replicated or audited, and where its conclusions feed health-technology assessment or regulatory decisions, non-verifiability becomes an evidence-base integrity risk. Farotimi et al. [22] adopt RAISE and DEST principles to mandate analysis code, open datasets, random-seed documentation, and explicit disclosure of model limitations; while Galli et al. [26] strictly mandate the disclosure of a verbatim prompt history, structured human-verification procedures, and explicit AI-versus-human stage declarations, Hasan et al. [88] require detailing the LLM version, interface, and timestamp of use, alongside the extent of human supervision, data entry methods, and a description of the iteratively refined prompts to the extent possible. The field's constraint is adoption, not invention: the tools and norms already exist but have not yet been operationalised as mandatory standards.

Table 10. Transparency and reproducibility practices in the T5 sub-corpus and the wider 122-paper corpus. Figures are approximate estimates from evidence mapping.

Practice	Description	Sub-Corpus (%)	Full Corpus (%)
Verbatim prompt disclosure	Exact prompt text reproduced in paper or supplement	∼15%	<10%
Model version specified	Exact model identifier and access date stated	∼60%	∼50%
Temperature/top-p disclosed	Inference parameters explicitly reported	<5%	<2%
Human verification of AI output	Explicit statement of human review of AI-generated content	∼90%	∼75%
PRISMA 2020 item 24 explicit	Non-financial support/AI use disclosed per item 24	∼25%	<10%
Code or reproducibility materials released	Code, prompts, or structured outputs publicly available	∼20%	∼15%

7. Agentic AI and End-to-End Systematic Review Pipelines

7.1. From AI-Assisted to Agentic Review

The foregoing sections characterise LLM performance on discrete, single-stage SLR tasks: query generation, screening, extraction, quality appraisal, synthesis. A distinct and more ambitious paradigm is the *agentic* systematic review, in which an AI system executes multiple review stages sequentially, passing outputs between stages, invoking specialist tools, and operating without continuous human intervention at each step. The distinction is not merely taxonomic: an agentic pipeline changes the economics of the review by reducing the marginal cost of each incremental stage, while concentrating risk by allowing errors from one stage to propagate downstream without detection.

The majority of multi-stage pipeline studies in the corpus are post-2024, consistent with the availability of long-context models and tool-use APIs in commercial LLM offerings (Section 3.2). A central observation is that none of the multi-stage pipeline papers in the corpus implements a formally multi-agent architecture—one with distinct autonomous communicating agent roles, graph-based orchestration, or standardised inter-agent interfaces such as the Model Context Protocol. Without exception, the pipeline papers describe single-agent LLMs with cascading logic, or sequential human-orchestrated workflows in which two or more tools are used in succession.

Liang et al. [50] deploy ChatGPT (via the Consensus plugin) for literature retrieval alongside DeepSeek for multi-round extraction and synthesis in a photovoltaic power plant site-selection review; however, their workflow is explicitly human-in-the-loop, requiring

researchers to manually save and transfer documents between tools rather than relying on automated inter-model communication. Noroozi et al. [116] construct a two-component architecture combining a Yake NLP-based search query refiner with a TF-IDF and Random Forest relevancy classifier, but the components similarly require human experts to feed keywords between stages, and the relevancy classifier requires human labelling of an initial training batch. These cases confirm the general pattern across the corpus: what is described as an agentic or automated pipeline is, in practice, a multi-tool workflow in which a human operator remains the integration layer between components.

Figure 12 and Table 11 classify the pipelines discussed in this review into four levels of automation, from manual (L0) to fully autonomous (L3).

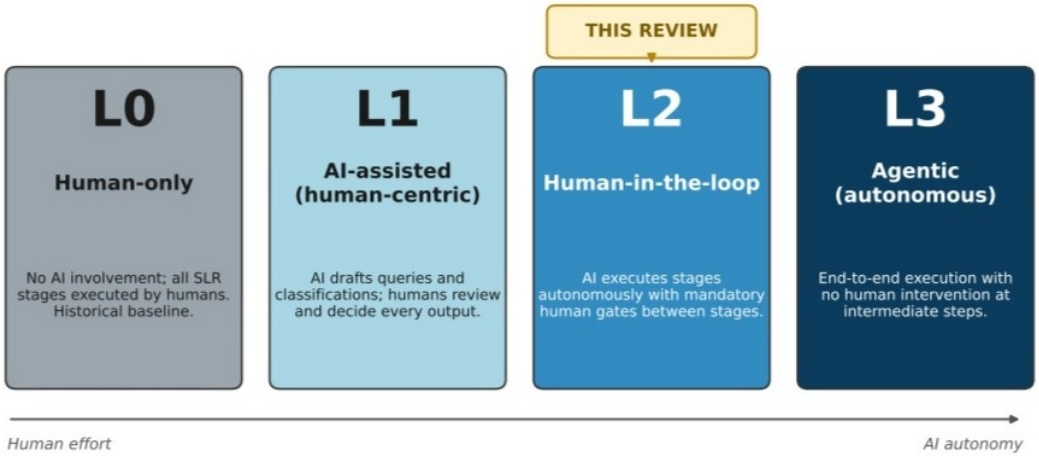


Figure 12. AI automation levels in systematic reviews (L0–L3). Human-to-AI effort ratio and representative tools or studies are shown per level (classification extended from the T5 corpus, $n = 36$; adapted from Köhler et al. [7]). No corpus paper fully instantiates L3 (fully autonomous); the present review's pipeline operates at L2 (human in the loop), with mandatory human decision gates at every stage—to our knowledge the first documented MCP-enabled, tool-orchestrated agentic-assisted L2 pipeline.

7.2. Agentic Pipeline Architectures in the Corpus

The multi-stage pipeline studies in the corpus exhibit three architectural patterns (Supplementary Table S3).

Single-agent cascading pipelines use one LLM across sequential stages. Torres et al. [2] (PROMPTHEUS) chain arXiv retrieval, Sentence-BERT filtering, BERTopic clustering, and GPT-4o synthesis into a LaTeX document, achieving high ROUGE-1 precision (0.969–0.995) but low recall (0.026–0.039). Purewal et al. [11] demonstrate a three-step pipeline achieving "near-perfect accuracy" on meta-analytic pooling despite only 70.4% accuracy at abstract screening. Alshami et al. [106] provide an early demonstration with GPT-3.5, achieving 88% screening accuracy in batches of 5–10 abstracts.

Multi-tool human-orchestrated workflows combine heterogeneous tools with human transfer between components. Scherbakov et al. [8] integrate Covidence, GPT-4o, Scite.ai, and NotebookLM across screening, extraction, and drafting. Liang et al. [50] combine Consensus for retrieval with DeepSeek for extraction, with researchers manually transferring documents between platforms.

Table 11. AI automation levels in systematic review pipelines (L0–L3), adapted from Köhler et al. [7]. In Köhler et al.'s corpus, human in the loop (L2) was the most frequently documented configuration (40.7%). No corpus paper in the present review fully instantiates L3; the present review's own pipeline operates at L2 (human in the loop), and is—to our knowledge—the first documented MCP-enabled, tool-orchestrated agentic-assisted L2 pipeline.

Level	Classification	Defining Characteristics	Representative Studies
L0	Human-only	No AI involvement; all SLR stages executed by human reviewers; constitutes the historical baseline	Pre-2022 conventional SLR practice; provides gold-standard comparators used throughout the corpus
L1	AI-assisted (human-centric)	AI generates first-draft outputs (queries, summaries, classifications); all outputs reviewed and accepted or rejected by humans before use; human makes every consequential decision	[39] (search terms); [106] (screening batch review); [49] (manual theme synthesis); [38] (query peer review)
L2	Human in the loop	AI executes defined pipeline stages autonomously but with mandatory human decision gates between stages; most frequent configuration in the extended LLM-SR literature (40.7% [7]); workload reduction 33–93% possible [18]	[2] (mandatory human review stage); [5] (two-reviewer validation); [50] (human document transfer between models); [11] (human FT stage); the present review (MCP-enabled pipeline with a mandatory human validation gate at every stage)
L3	Agentic (autonomous)	AI system executes all pipeline stages end-to-end without human intervention at intermediate steps; invokes specialist tools; maintains persistent state across sessions; not yet demonstrated in any peer-reviewed corpus paper [10,107]	None; not demonstrated in any peer-reviewed corpus paper, nor in the present review (which operates at L2)

Stage-level mapping studies characterise the automation landscape, confirming the stage asymmetry documented in Section 1.1: quality assessment remains largely untouched despite substantial investment in screening automation. Soprano et al. [25] explicitly map the stages of their AI-aided framework directly to corresponding PRISMA 2020 checklist items.

7.3. Failure Modes of Agentic Pipelines

Single-stage failure modes—hallucination (Section 6.2), calibration errors (Section 4.2), context sensitivity—are amplified in agentic pipelines because outputs from one stage become inputs to the next: an error in screening that silently excludes a relevant study means that study is absent from extraction and synthesis, with no mechanism for downstream detection.

Error propagation between stages. Chelstrom et al. [56] document that applying manual evaluation only after automatic LLM classification creates a risk that relevant studies are permanently excluded "without the reviewer's knowledge." Unlike single-stage screening errors—which can be detected by sampling excluded records—pipeline errors propagate silently unless an explicit verification gate is inserted between stages.

Figure 13 illustrates the compounding effect: without human oversight gates, AI-only synthesis tools achieve primary-study recall rates as low as 38.1% [111], meaning more than 60% of relevant studies may be lost silently before any synthesis output is produced.

Context-window limitations. Context-window constraints are a systematically documented architectural failure mode: Bayani et al. [13] cap at 4098 tokens and exclude references and results sections; Surles and Noteboom [129] report hallucination and misrepresentation when more than five papers are included in a session; and Felizardo et al. [14] observe that while token limits were not a barrier when screening the title and abstract of one study at a time, the context window parameter cannot be ignored when reviews progress to the full-text evaluation stage. Context-window management is therefore a non-trivial engineering requirement for any pipeline processing full-text documents.

Hallucinated citations. Schopow et al. [16] document that in a composition-phase SLR pipeline, ChatGPT fabricated 3 of 28 supplied references, supplied 5 thematically pertinent but non-substantiating references, and required corrections to publication years on 2 more; hallucinated citations require independent bibliography verification of every cited work because they evade field-level accuracy checks.

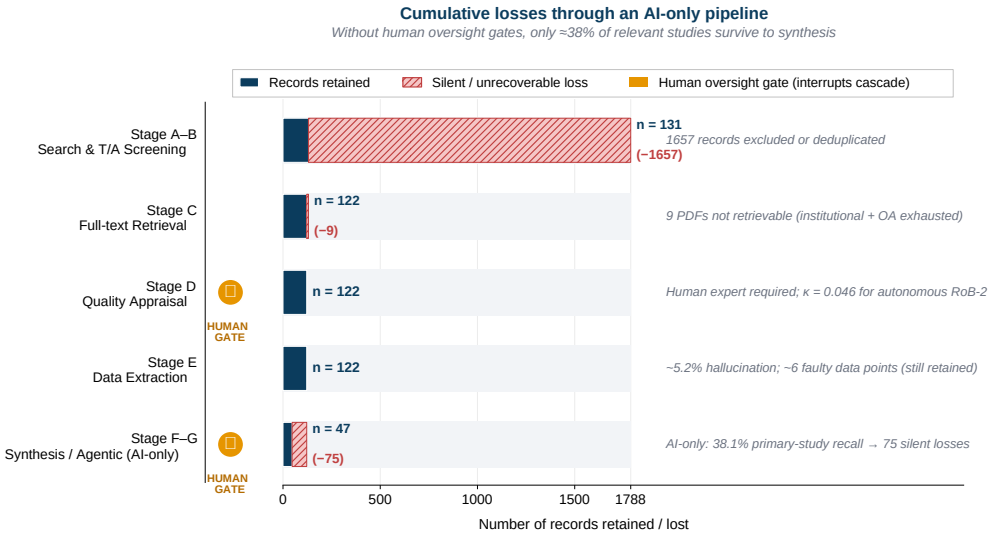


Figure 13. Cumulative error propagation in an AI-only systematic review pipeline. At each stage, silent failures reduce the effective study pool without the downstream pipeline detecting the loss. Human oversight gates (orange markers) interrupt the cascade. The synthesis stage (F–G) shows projected losses based on the 38.1% primary-study recall reported by Tosi [111] for a GPT-4-powered AI-only literature review tool; these values are illustrative of comparable AI-only pipelines and do not represent observed losses in this review's own pipeline.

7.4. Reference Architectures and Deployment Requirements

The agentic pipeline literature has not converged on a canonical reference architecture, but the stage-specific tractability profile established in Section 7.2 has direct implications for hybrid design: stages amenable to automation (screening, search) can be automated with established techniques, while quality assessment requires human expertise. Kusa et al. [109] propose a conceptual evaluation framework for benchmarking LLMs across the full pipeline without training-data contamination, establishing evaluation criteria rather than reporting performance. Torres et al. [2] (arXiv-indexed domains) and Purewal et al. [11] (predefined clinical questions) demonstrate feasibility for constrained tasks but omit quality appraisal and at-scale full-text processing; both are insufficient for end-to-end PRISMA compliance. Evaluations of ten commercial tools [107] and of six LLMs across three sequential SLR tasks [10] converge on a task-specific superiority pattern (no single family dominates), with all six models in Sollini et al. [10] failing manuscript drafting; a meta-analysis in orthodontics [130] similarly concludes that expert oversight remains essential. Collectively with the near-absence of quality-appraisal automation (Section 7.2) and silent error propagation (Section 7.3), these findings indicate that current agentic pipelines are research prototypes rather than production-ready systems. The stage-specific validation requirements derived from this evidence are summarised in the Supplementary Materials (Table S4).

8. Cross-Cutting Synthesis Matrix

The per-stage analyses in Sections 4–7 reveal seven challenge dimensions that recur across all PRISMA-defined stages and that define the shared boundaries of current LLM capability in systematic reviewing: calibration and confidence misestimation; hallucination

and value fabrication; human–AI inter-rater agreement; reproducibility and versioning; domain and language generalisation; context-window management; and error propagation in pipelines. Figure 14 summarises their severity across the seven stages; the full cross-cutting evidence map—best current approach and remaining gap per challenge—is provided in Appendix C.

Severity was assigned on a reproducible four-point ordinal scale applied to each challenge–stage cell: **0 (none)**—no corpus evidence that the challenge materially affects the stage; **1 (minor)**—the challenge is documented but current practices mitigate it, with residual impact on efficiency rather than on validity; **2 (moderate)**—the challenge measurably degrades output quality and requires active human mitigation; **3 (severe)**—the challenge threatens the validity of the stage's output, and no corpus-documented method reduces it to acceptable levels. Each cell was scored from the stage-specific evidence synthesised in Sections 4–7 and mapped to the corresponding row of the cross-cutting evidence map (Appendix C), which records, per challenge, the best current approach and the remaining gap that determined the score; where studies disagreed, the more conservative (higher) score was assigned. The resulting matrix is therefore an evidence-anchored ordinal synthesis rather than a quantitative measurement.

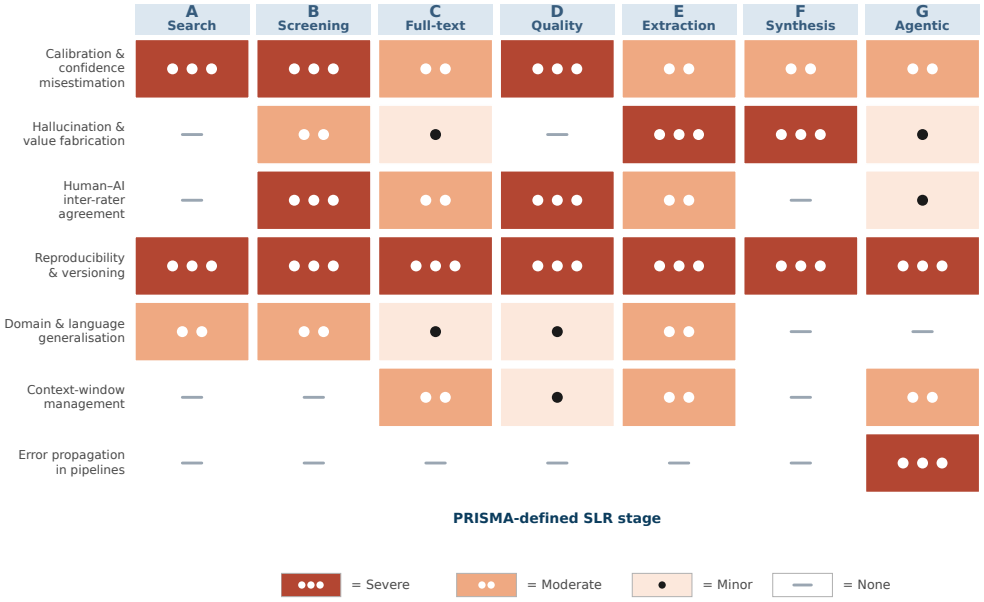


Figure 14. Severity of cross-cutting challenge dimensions across PRISMA-defined SLR stages (0 = no issue; 3 = severe). Calibration and human–AI agreement challenges are most acute at Stage B (Screening) and Stage D (Quality Appraisal); error propagation is unique to Stage G (Agentic pipelines). Scores follow the four-point rubric defined in the text and are anchored to the per-challenge evidence in Appendix C.

The concentration of high-severity cells in Stages B and D confirms screening calibration and quality-appraisal agreement as the two most consequential bottlenecks across the review lifecycle. Reading the evidence map two tiers emerge. Tier 1 comprises challenges for which partial solutions exist, and the remaining gap is standardisation rather than capability—threshold calibration (Section 4.2), reproducibility via model-version disclosure and fixed random seeds (Section 6.3)—and could be operationalised in near-term deployment guidelines. Tier 2 comprises open problems for which no current mitigation reduces risk to acceptable levels: RoB calibration (Section 5.3), automatic hallucination detection in synthesis prose (Section 6.2), cross-stage error-propagation prevention (Section 7), and full reporting transparency (Section 6.3). Quality appraisal (Stage D) is where the

gap between current capability and production readiness is widest and most resistant to prompt-engineering remediation, because the required knowledge is procedural and domain-expert rather than pattern-matching—a type of knowledge LLMs approximate but do not master reliably.

9. Discussion

9.1. Principal Methodological Debates in the Corpus

The corpus supports four active methodological debates, each reflecting a genuine empirical disagreement rather than a terminological or definitional dispute. Stage-level metrics that support the following summaries are reported in Sections 4–7 and are not restated here.

Prompt engineering versus fine-tuning for search and screening. Adam et al. [37] report an ≈73 pp sensitivity gap between a fine-tuned Mistral-Instruct-7b model and zero-shot GPT-4 for query generation, but at a precision cost that inflates NNR to 908–1206. Against this, structured prompt engineering with explicit in/exclusion criteria [26,71] and QA-framework prompts on commodity LLMs [51] recover competitive performance without task-specific training. The practical resolution is likely scale-dependent: fine-tuning for large single-domain programmes, prompt engineering for heterogeneous reviews at smaller teams.

GPT-family versus open-source alternatives. The corpus is GPT-dominated (49/122 studies), reflecting historical API availability rather than proven superiority. Task-specific superiority varies across families [10,13], and open-source models can outperform GPT-4o-mini in screening sensitivity, albeit with slightly lower specificity [62]. A 4609-study analysis of clinical AI evaluations concludes that conclusions about model performance must be "interpreted cautiously", given that only 19 qualified as prospective RCTs [86].

Single-reviewer replacement versus dual-reviewer assistance. This is the most practically consequential debate. Evidence on replacement is genuinely contested: some studies argue the second human reviewer is needed only to adjudicate disagreement [18], whereas others find that LLMs are not reliable enough to substitute for human reviewers as independent screeners [63]; Zrubka et al. [77] explicitly recommend LLMs as a supplementary tool alongside two human raters to reduce both false negatives and false positives. Trad et al. [1] report a 0% false-negative rate in a RAG-grounded pipeline—one of the strongest accuracy results in the corpus. The debate is the most directly relevant to adoption decisions of health technology assessment bodies.

Degree of agentic autonomy. Fully automated pipelines achieve high precision but recall too low for standalone use [2,111]; human-in-the-loop architectures recover substantial efficiency without compromising recall [5,107], and remain the most frequent configuration across the corpus (Section 7.2). Current agentic systems therefore remain substantially incomplete relative to human-authored reviews.

Figure 15 provides a high-level orientation to the four-RQ structure of the findings: RQ1 and RQ2 yield deployable conclusions, while RQ3 and RQ4 delineate the infrastructure investments required to close outstanding gaps. The key evidence based on the T1-T5 sub-corpora for RQ2 includes [1,32,117,123], and for RQ3 [22,26,88] from the T5 sub-corpus.

Table 12 synthesises the four principal debates into a cross-cutting summary aligned to the four research questions of this review.



Figure 15. Summary of findings across the four research questions. Each quadrant presents the principal finding, key evidence, and resolution status: RQ1 and RQ2 are substantially resolved; RQ3 and RQ4 identify critical infrastructure gaps and a structured research agenda.

Table 12. Cross-cutting synthesis of findings by research question. Quantitative evidence and citations are in Sections 4–8; performance tables are Supplementary Table S6 and Tables 9 and 10.

RQ	Principal Finding	Strongest Evidence	Critical Limitation	Implication
RQ1	Performance is task-specific; sensitivity/specificity asymmetry is consistent across screening; quality appraisal shows near-chance agreement.	Screening sens. 85–100% [17]; extraction accuracy 96.3% [31].	As established in Section 5.3, RoB instruments yield near-chance agreement with central-tendency bias documented across all instruments.	LLMs suitable as high-recall pre-screeners; quality appraisal requires mandatory expert review until $\kappa > 0.60$ achieved.
RQ2	No published pipeline endorses full automation; human in the loop (40.7%) dominates [7]; efficiency gains are large but stage-contingent.	≈99% processing-time reduction [5]; 33–93% workload reduction in dual-reviewer designs [18].	Unguided search recall 12.9% vs. ∼74% human [42]; fine-tuned model NNR 908–1206 vs. 580 for human queries [37].	Adopt L2 human-in-the-loop design with human decision gates after each automated stage; no stage currently supports fully autonomous operation.
RQ3	Hallucination rates are stage-specific and quantifiable; systematic biases in appraisal are structural, not random.	Extraction hallucination 2.3–5.2% [32,123]; reference hallucination 28.6–91.4% [108]; PRISMA adherence overestimated by 23.1–29.7% [105].	No automatic hallucination-detection method with acceptable recall demonstrated at synthesis stage [102].	Source-linked extraction, schema-constrained JSON output, and an inviolable human post-editing gate are mandatory at synthesis stage.
RQ4	Clinical science dominates (≈34% of corpus); performance is domain-structure-dependent rather than domain-specific.	CONSORT compliance F1 = 0.86 [36]; structured-domain performance generally higher than unstructured narrative tasks.	No cross-domain meta-regression published; law, economics, and humanities absent; only 19/4609 clinical AI studies qualify as prospective RCTs [86].	Findings should not be extrapolated across domains without domain-specific validation; structured review stages generalise better than interpretive ones.

9.2. Limitations of This Review

This review has several limitations that affect the generalisability of its conclusions.

Database and source coverage. The search was restricted to three databases (Scopus, PubMed, IEEE Xplore) and to peer-reviewed sources. Web of Science, the ACM Digital Library, and DBLP were not searched separately; although Scopus indexes much of their content, venues indexed only in ACM DL or DBLP may have been missed. Google Scholar was not used as a primary source because its results are not reproducibly exportable or deduplicable and include substantial non-peer-reviewed and duplicate material, which is incompatible with a PRISMA-reproducible protocol. Preprint servers (notably arXiv) were excluded, so the newest preprint-stage agentic pipelines—many first appearing in 2025–26—are under-represented; a preprint-inclusive search is a priority for future updates.

Single-team screening without a formal inter-rater statistic. Human screening decisions were made by author-team consensus with adjudication of all uncertain records rather than by dual independent screening; a formal inter-rater agreement statistic (e.g., Cohen's κ) was therefore not computed. This is a deviation from the dual-reviewer ideal of PRISMA 2020, mitigated by the recall check reported in Section 2.5 (no eligible record was missed in the audited sample).

Pace of field and rapid model obsolescence. The corpus spans 2022–March 2026, during which the underlying LLM technology changed substantially with each major release. Performance benchmarks obtained on GPT-3.5 should not be used to justify or reject deployment of GPT-4o; findings from GPT-4 evaluations may not transfer to Claude 3.7 or GPT-4o. Studies published in 2022–2023 describe a technological context qualitatively different from the 2025–26 landscape, and aggregate findings necessarily reflect this heterogeneity. Non-determinism compounds obsolescence: Wang et al. [42] document a query recall variance of 12% of the mean value across ten identical runs of the same prompt, demonstrating that even a correctly specified benchmark may not yield reproducible results across repeated queries.

Benchmark incomparability. The absence of a shared benchmark across studies means performance claims are not directly comparable: studies report accuracy, recall, F1, Cohen's κ, or agreement rate against review-specific gold standards with incompatible reporting conventions. Several papers use the CLEF TAR benchmark, enabling limited direct comparison, but the majority use proprietary or study-specific reference sets that cannot be replicated.

Language restriction. The corpus includes only English-language studies, excluding substantial work in Chinese, Japanese, and other languages. Bourgeois and Ellingson [38] document an additional bias: in two of their three AI trials, the model inserted an unprompted English-only concept filter into the generated search string, thereby restricting the retrieval set to English-language sources despite this parameter never being requested—an LLM-introduced limitation that may compound the human decision to restrict corpus eligibility to English.

Publication bias toward positive results. Studies demonstrating LLM failure at a specific SLR stage are less likely to be submitted or accepted than those demonstrating partial success. Chen et al. [86] characterise this pattern quantitatively in the clinical AI literature: of 4609 studies reviewed, only 19 qualified as prospective RCTs, and at least 25% used fewer than 30 instances, demonstrating a strong imbalance toward simulated and synthetic data over real-world clinical evidence.

AI-assisted synthesis of an AI-assisted-review corpus. The use of Claude Sonnet 4.6 and NotebookLM to assist the synthesis of a corpus evaluating AI-assisted reviews creates a circularity acknowledged in Section 2.11. The present review's own AI transparency practices exceed the median corpus practice, but they were shaped in part by reading the same corpus papers—a recursive dependency that cannot be fully disentangled.

9.3. Threats to Agentic Review Validity

The evidence in Section 7.3 and 8 defines five principal threats that deploying organisations must address. *Cross-stage failure cascades:* classification errors propagate silently [56], requiring inter-stage sampling of excluded records. *Context-window truncation:* multi-study full-text processing remains problematic when full-text evaluation is required, as context-window constraints become critical and cannot be ignored [14]. *Confabulated citations:* any pipeline generating bibliography entries must cross-check outputs against a verified corpus [16]. *Eligibility-criteria drift:* session-to-session stochasticity shifts extraction error locations even when aggregate accuracy is stable [31], requiring a persistent machine-readable criteria representation in every prompt. *Quality-appraisal shortfall:* LLMs exhibit low agreement levels on validated risk-of-bias instruments (Section 5.3), indicating that automated quality assessments currently require responsible human supervision and should be combined with expert judgment [35,87].

10. Conclusions and Research Agenda

10.1. Principal Conclusions

This systematic review of 122 primary studies published between 2022 and March 2026 supports six conclusions, each grounded in the evidence synthesised in Sections 4–8; stage-level metrics are not restated here.

(1) Title/abstract screening is the most mature stage for LLM automation, but human oversight remains mandatory (Section 4.2). High-sensitivity configurations [17] trade off against low precision; open-source models can outperform proprietary GPT-4o-mini in screening sensitivity, albeit with slightly lower specificity [62]; no configuration achieves simultaneously $\geq$95% recall, acceptable precision, and cross-session reproducibility without a human checkpoint.

(2) Structured data extraction is conditionally capable, with field-stratified human oversight required (Section 5.1). Hallucination rates of 2–5% concentrate on outcome effect estimates and fields absent from the source text [32,123]; schema-constrained JSON output with zero-temperature settings [33,99] and hybrid dual-extraction architectures [125] substantially reduce error rates. Meta-analytic pooling is the one demonstrably reliable synthesis sub-task (Section 6.1).

(3) Quality appraisal remains the least automated and least reliable stage (Section 5.3): the best documented LLM–human agreement is $\kappa = 0.0455$ for RoB 2 [87] and $\kappa = 0.297$ for ROBIS [35], with structural central-tendency bias; all decisions must currently route through human experts.

(4) Evidence synthesis and reporting require an inviolable human post-editing gate (Section 6.2). No corpus study reports AI-generated synthesis prose published without human revision [49,106]; fabricated synthesis claims that cite real papers can be undetectable without re-reading every source [102].

(5) Agentic pipelines are technically feasible but not production-ready (Section 7). No corpus study satisfies all five production-readiness requirements—persistent protocol specification, RAG corpus grounding, inter-stage verification, context-window management, and a transparent audit trail—and the human-in-the-loop approach remains the most frequent configuration in the extended LLM-SR literature (40.7% [7]).

(6) The reporting transparency gap is critical and addressable (Section 6.3). Temperature is disclosed in fewer than 2% of corpus papers, verbatim prompts in fewer than 10%, and PRISMA 2020 item 24 mapping in fewer than 10% (Table 10); the strategies required—exact model-version disclosure, verbatim prompt release [49], and CONSORT-AI-aligned reporting [26]—already exist. The constraint is adoption, not invention.

The present review is itself an attempt to satisfy all six conditions it identifies. The agentic pipeline used to conduct it—MCP-connected database search, multi-stage hybrid screening, retrieval-augmented synthesis grounded exclusively in the uploaded corpus, and a two-stage citation audit applied to every in-text claim—addresses each of the five production-readiness requirements stated in conclusion (5) and implements each of the transparency practices called for in conclusion (6). It does not do so perfectly: quality appraisal (conclusion 3) remained fully human-executed throughout, and the proprietary nature of NotebookLM introduces a non-reproducibility risk that is explicitly acknowledged in Section 2.11. The methodology therefore constitutes an early, human-supervised instantiation of the reference architecture proposed in Section 10.2 rather than a finished production system—but one that demonstrates the architecture is operationally coherent and executable today.

10.2. A Reference Architecture for Agentic Systematic Reviews

Drawing on the convergent requirements identified in Section 7.4 and the gap analysis in Section 8, Table 13 proposes a seven-layer reference architecture for a production-ready agentic systematic review system. The architecture is not purely theoretical: as documented in Section 2.12, the methodology of this review constitutes an early proof-of-concept instantiation of all seven layers, conducted under continuous human oversight and subject to the verification standards each layer prescribes.

Table 13. Seven-layer reference architecture for agentic systematic reviews. Each layer specifies its function, key technique, and corpus evidence base.

L	Layer	Function	Key Technique	Corpus Evidence
1	Protocol	Machine-readable, version-controlled PCC/PICO criteria injected into every downstream prompt	Structured format preventing eligibility-criteria drift	[90]
2	Search	LLM-assisted synonym expansion and database-specific syntax translation	Mandatory human information-specialist sign-off	[39]
3	Screening	Multi-stage triage: CoT for T/A, three-way classifier for FT	Seed-study calibration; 5–10% excluded-record sampling	[5,56,77]
4	Extraction	Schema-constrained extraction with JSON-mode and zero-temperature	Field-stratified human spot-checking; RAG corpus grounding	[1,33]
5	Appraisal	Full human quality appraisal; LLM pre-populates instrument fields as draft	Human review mandatory until $\kappa > 0.60$ demonstrated	[35]
6	Synthesis	RAG-grounded synthesis drafts with citation verification	Mandatory human post-editing gate	[32,123]
7	Transparency	Audit trail linking every output to prompt, model version, temperature, source	Verbatim prompt release; PRISMA item 24 disclosure	[22,26,88]

Figure 16 presents the reference architecture as a layered stack, making explicit which layers permit progressive automation (Layers 1–4, 6–7) and which require human expert execution as a non-negotiable condition (Layer 5, Quality Appraisal).

The architecture should be read as an evidence-derived *reference framework* rather than an independently benchmarked system, and we make its validation status explicit. Each layer carries one of three evidential statuses (Table 14): *empirically supported*, where the layer's key technique is backed by quantitative results in the corpus; *partially supported*, where corpus evidence establishes feasibility but not production-grade reliability; and *conceptual/normative*, where the layer is justified by convergent argument and this review's own proof-of-concept but is not yet validated against an external gold standard. We apply

three validation criteria: (i) each layer's key technique is traceable to at least one corpus study reporting a quantitative outcome; (ii) the layer's prescribed human-oversight gate is justified by a documented failure mode; and (iii) the end-to-end configuration is positioned against representative published pipelines (Table 15). By these criteria, Layers 2–4 are empirically supported, Layers 1 and 6 are partially supported, and Layers 5 and 7 are presently conceptual or normative; full independent benchmarking against alternative architectures on a shared task remains future work (Section 10.3).

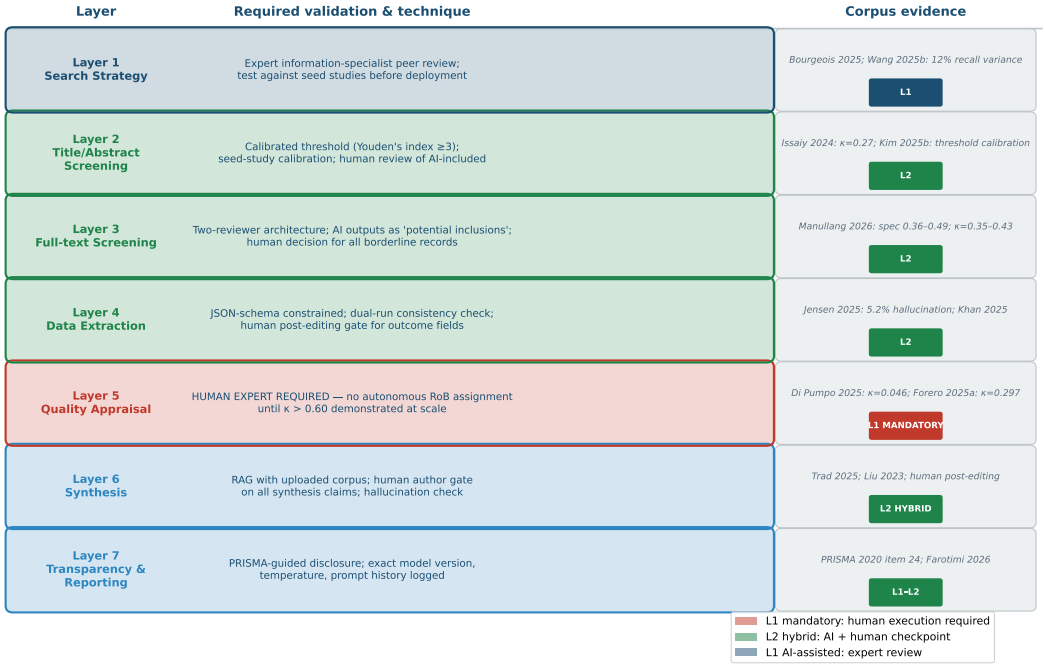


Figure 16. Proposed seven-layer reference architecture for production-ready agentic systematic reviewing. Layers 1–4 are amenable to automation levels L1–L2 with appropriate human checkpoints [29,32,38,63,112]; Layer 5 (Quality Appraisal) requires full human execution until inter-rater agreement exceeds $\kappa = 0.60$ [87]; Layer 6 also requires human author gating on all synthesis claims to avoid hallucinations [1,102]; Layer 7 mandates PRISMA-guided disclosure [19] of all AI-generated content.

To position the proposed configuration against existing practice, Table 15 compares this review's L2 instantiation with three representative published pipelines from the corpus, spanning the autonomy spectrum from near-autonomous to human in the loop.

Table 14. Validation status of each architectural layer against the corpus evidence base.

L	Layer	Status	Basis in Corpus
1	Protocol	Partially supported	Criteria-drift prevention demonstrated [90]; no controlled comparison of drift with/without machine-readable criteria.
2	Search	Empirically supported	LLM query expansion with mandatory information-specialist sign-off; unguided recall is inadequate without it [39,42].
3	Screening	Empirically supported	High-sensitivity screening across >20 configurations; abstract-screening $\kappa > 0.89$, F1 0.950 [5,56,77,124].
4	Extraction	Empirically supported	High field-level accuracy—96.3%, F1 0.98 [31]; F1 0.974 at scale (117,889 data points) [33]—but falls to ∼80% on primary-outcome effect estimates and to 16–35% on complex fields, so field-stratified human spot-checking with RAG grounding is required [1,31,125].
5	Appraisal	Conceptual (human-mandated)	Human–LLM agreement is near zero on validated risk-of-bias instruments (overall RoB-2 $\kappa = 0.044$; 27.8% agreement) [87], with intra-rater ICC 0.518 and systematic overestimation of reporting quality [35,89,105]; no source endorses replacing human appraisers, so the layer prescribes full human execution.

Table 14. *Cont.*

L	Layer	Status	Basis in Corpus
6	Synthesis	Partially supported	RAG-grounded synthesis is feasible (BERTScore F1 0.88 [99]; tool-synthesised evidence preferred over baseline GPT-4 in 62.5–100% of cases [43]), but unguarded LLM drafting fabricates unsupported claims [102] and biases risk-of-bias output [100], mandating a human post-editing gate.
7	Transparency	Normative	Disclosure practices proposed from convergent reporting-gap evidence; not benchmarked [22,26].

Table 15. Comparison of the proposed architecture with representative corpus pipelines.

Pipeline	Autonomy	Layers Covered	Reported Outcome/Validation
This review	L2 (HITL)	1–7 (all, PoC)	Proof-of-concept instantiation with human gates at every stage; two-stage citation audit. Framework not yet externally benchmarked.
Torres et al. [2]	Near-autonomous	Search–Synthesis	High precision, but recall too low for standalone use; study-specific evaluation.
Tripathi et al. [5]	L2 (HITL)	Screening–Extraction	≈99% processing-time reduction with recall preserved; study-specific evaluation.
Trad et al. [1]	L2 (RAG, HITL)	Screening	0% false-negative rate in a RAG-grounded pipeline; study-specific evaluation.

10.3. Open Research Gaps and Agenda

Figure 17 reveals that G1, G4, and G7 are structural prerequisites: resolving them in parallel would unlock progress on the remaining five gaps simultaneously and make G8 (full prospective validation) tractable.

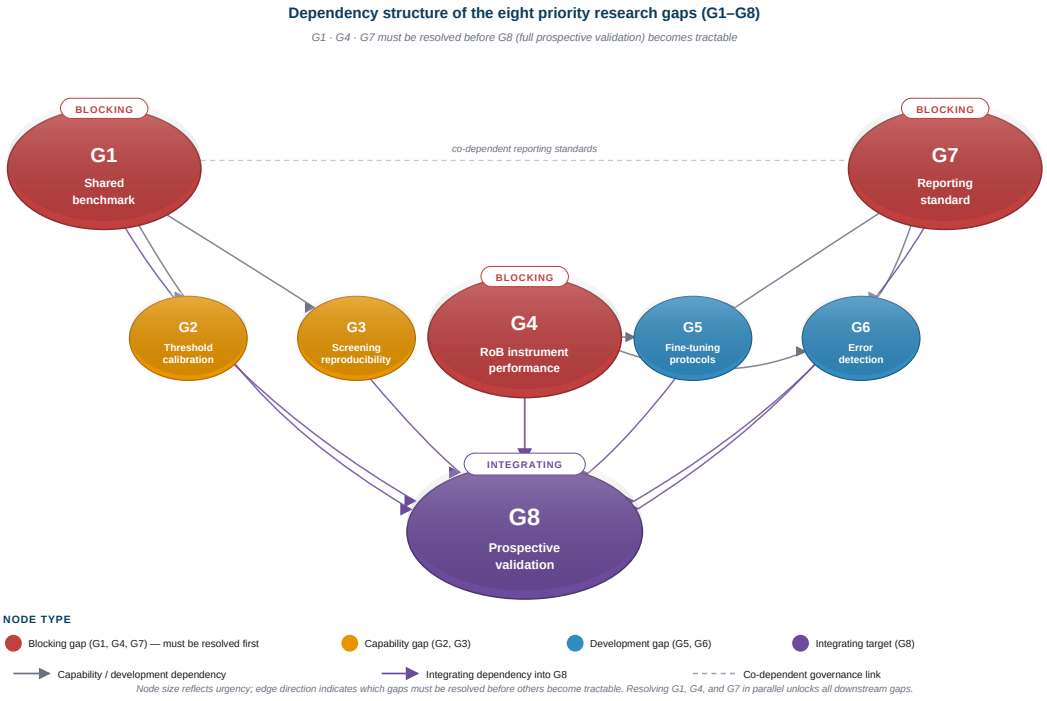


Figure 17. Dependency structure of the eight priority research gaps (G1–G8). Node size reflects urgency; directed edges indicate which gaps must be resolved before others become tractable. G1 (shared benchmark), G4 (RoB instrument), and G7 (reporting standard) are structural blockers that constrain progress on all downstream gaps; G8 (prospective validation) integrates all prior investments.

The eight gaps in Table 16 collectively define the research and standards agenda required to close the distance between the current state of LLM-assisted systematic reviewing and the validated, general-purpose agentic system that the field requires. The most urgent

priorities—G1 (shared benchmark), G4 (RoB instrument performance), and G7 (reporting standard)—are infrastructure investments that would accelerate progress on all other gaps, because they establish the measurement and reporting standards against which subsequent capability claims can be evaluated. Until G7 is adopted uniformly by journal and editorial bodies, the evidence base for LLM capability in systematic reviewing will remain partially non-verifiable regardless of the technical advances being made in underlying LLM technology (Section 6.3). Until G4 is solved, any agentic pipeline achieving acceptable screening and extraction recall must still route every included study through a human appraiser before the evidence base can be considered GRADE-compliant (Section 5.3).

Table 16. Open research gaps in LLM-assisted and agentic systematic reviews. Evidence column cites the specific corpus findings that define the gap.

#	Gap	Stage(s)	Recommended Action	Defining Evidence
G1	Shared multi-domain benchmark for screening across model generations	B	Establish and maintain a publicly versioned benchmark (building on CLEF TAR) covering ≥10 domains with model-version tracking	Performance varies from $\kappa = 0.186$ to $\kappa = 0.895$ depending on domain [35,53]; no cross-domain benchmark exists
G2	Standardised inter-rater agreement threshold for LLM–human screening	B	PRISMA extension working group to specify minimum acceptable κ for LLM screening certification, with domain stratification	No agreed κ threshold; 100% sensitivity at 14% precision is within current range [17]; workload reduction 33–93% [18]
G3	Field-stratified hallucination detection for structured extraction	C, E	Develop an automated hallucination-detection module cross-referencing extracted values against the source PDF at field level	5.2% hallucination rate for absent data [32]; errors shift location across sessions even when overall accuracy is stable [31]
G4	LLM performance on validated RoB instruments above $\kappa = 0.60$	D	Domain-specific fine-tuning on ROBINS-I, RoB 2, AMSTAR 2 with supplementary protocol documents in context; evaluation of long-context models for Domain 5 resolution	Best ROBIS $\kappa = 0.297$ [35]; RoB 2 $\kappa = 0.0455$ [87]; structural central-tendency bias documented in both
G5	Automatic hallucination detection in synthesis prose	F	RAG-grounded synthesis with automatic claim-to-source entailment verification; evaluation against human expert annotation of claim veracity	No automatic detection with acceptable recall demonstrated; fabricated synthesis claims may be undetectable without re-reading all sources [32,102]
G6	Cross-stage error-propagation quantification in agentic pipelines	G	Fault-injection studies measuring downstream impact of deliberate Stage B and E errors on final synthesis conclusions	T/A sensitivity 54.9% propagates to exclude ∼half of relevant literature [11]; ROUGE-1 recall 0.015–0.039 for compiled documents [2]; AI-only pipeline recall 38.1% [111]
G7	Minimum reporting standard for AI-assisted systematic reviews	F (all)	PRISMA extension or dedicated reporting guideline mandating model version, prompts, parameters, verification procedure, and AI-contribution statement per stage	Temperature undisclosed >98% of corpus; verbatim prompts in <10%; PRISMA item 24 mapping in <10% [5,26,88]
G8	Validated agentic review pipeline for a real-world review	G	Conduct and publish a fully documented prospective systematic review using the reference architecture in Section 10.2, satisfying all five production-readiness requirements, with PRISMA 2020 fully compliant	No corpus paper satisfies all five requirements simultaneously [7,107]; human in the loop remains the most frequent configuration in the LLM-SR literature (40.7% [7])

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/informatics13100160/s1. Table S1: Per-study screening performance across the T1 sub-corpus; Table S2: Studies excluded at full-text review; Table S3: Representative multi-stage LLM-assisted systematic review pipelines in the corpus; Table S4: Stage-specific validation

requirements for AI-assisted systematic review pipelines; Table S5: LLM-generated search strategy performance across the T2 sub-corpus; Table S6: Key LLM data extraction studies; Table S7: Citation-audit results by manuscript section; Table S8: Twelve most significant citation corrections; Section S9: Full NotebookLM prompt/query set. The underlying data files (corpus list, MMAT appraisal workbook, citation-audit log, search strategy, review protocol and prompts) are openly available at Zenodo (DOI 10.5281/zenodo.22691555).

Author Contributions: Conceptualisation, E.A.M.-C.; methodology, E.A.M.-C.; software, E.A.M.-C.; validation, E.A.M.-C., I.Y., S.S., I.G. and D.P.; formal analysis, E.A.M.-C.; investigation, E.A.M.-C.; data curation, E.A.M.-C.; writing—original draft preparation, E.A.M.-C.; writing—review and editing, E.A.M.-C., I.Y., S.S., I.G. and D.P.; visualisation, E.A.M.-C.; supervision, I.Y. and D.P.; project administration, E.A.M.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data supporting the findings of this study are openly available in a citable Zenodo repository (DOI 10.5281/zenodo.22691555, CC-BY 4.0): the 122-study corpus with thematic tags, the full Boolean search strings, the per-study MMAT quality-appraisal workbook, the two-stage citation-audit log (all 612 key-slots), the prompt set, and the a posteriori review protocol. The full texts of the primary studies are available from their respective publishers. All notebook query responses were saved verbatim to time-stamped audit-evidence files before drafting commenced. Source notebooks are available for inspection upon reasonable request to the corresponding author; however, the proprietary nature of NotebookLM means that Google may modify or remove the service at any point, introducing an irreducible non-reproducibility risk that is acknowledged here explicitly.

Acknowledgments: During the preparation of this manuscript, the authors used large language model-based tools as part of the agentic review pipeline described in the Methods, for the purposes of search-query generation, title/abstract and full-text screening assistance, structured data extraction, and draft synthesis. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. No author has a financial relationship with any LLM vendor or systematic review tool provider evaluated in this study.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
GenAI	Generative Artificial Intelligence
SLR	Systematic Literature Review
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RoB	Risk of Bias
RQ	Research Question

Appendix A. NotebookLM Query Log (Representative Sample)

This appendix reproduces the standard grounding preamble and two representative queries. To keep the main text concise, the main prompt set has been moved to the Supplementary Materials (S9). The complete query–response evidence logs (totalling approximately 7700 lines, covering all ten manuscript sections together with the phase-0 corpus and sanity checks) are openly deposited (Zenodo, DOI 10.5281/zenodo.22691555).

Every analytic query began with the standard discipline preamble: *"Based ONLY on the sources in this notebook, reading each paper in its entirety—introduction, methods, results, discussion, and conclusions—not just the abstract or conclusions, . . . "* abbreviated as `[PREAMBLE]`

below. Correction queries followed the correction preamble described in the Methods (Section 2.11).

Representative Queries

Query A.1—corpus orientation (submitted to each notebook, T1–T6):

"What sources are loaded in this notebook? List every paper by title and, where available, author(s) and year. How many sources are there in total? Be exhaustive—do not omit any."

Query A.2—evidence synthesis (T3, Data extraction):

[PREAMBLE] What data extraction methodologies, schemas, and accuracy metrics are reported across these papers? How do they compare LLM performance to human reviewers? For each finding give the paper title, authors/year, the specific claim, and any methodological context (sample size, dataset, model). Use (Author, year) citations and be exhaustive.

Appendix B. Citation Audit Log Summary

Appendix B.1. Overview

A systematic citation-accuracy audit was conducted on all 612 in-text citation key-slots appearing in Sections 1–10 and Appendix C of this review. The audit used the six thematic NotebookLM notebooks (T1–T6, Appendix A) as the exclusive verification ground truth; no claim was classified from memory or general knowledge. For every cited claim, a verification query was submitted to the notebook containing the cited paper; the agent recorded one of four verdicts (below), and a human reviewer then recorded an independent resolution status for every key-slot. Where a claim was inaccurate or over-attributed, a correction was drafted from the notebook's own response, applied to the manuscript, and the passage re-audited before the section was closed.

Appendix B.2. Agent Verdict Labels

Acc. Accurate: The claimed fact, figure, or characterisation is directly and fully supported by the cited paper as confirmed by the notebook query.
Part. Partially Accurate: The substance of the claim is broadly correct, but one or more specifics (numerical value, exact phrasing, scope, or attribution of emphasis) deviate from the source text.
Fix Needs Correction: The claim is factually incorrect, materially mis-characterised, or incorrectly attributed to the cited key.
N/V **Cannot Verify:** The citation key does not correspond to one of the 122 corpus papers—it is a complementary reference introduced by the authors for additional context (e.g., a guidance document or methodological statement). Such claims were verified by the human team against the complementary source directly.

Appendix B.3. Aggregate Results

Supplementary Table S7 presents the verdict and resolution counts for the complete audit across all manuscript sections and Appendix C.

Appendix B.4. Key Findings

Of the 612 key-slots, the agent returned *Accurate* for 535 (87.4%), *Partially Accurate* for 35 (5.7%), *Needs Correction* for 20 (3.3%), and *Cannot Verify* for 22 (3.6%). Human validation confirmed all 535 Accurate slots, resolved the 55 Partially-Accurate and Needs-Correction slots (36 corrections applied to the manuscript), and accepted the 22 Cannot-Verify slots after checking the complementary reference each cited. The corrections were concentrated in the Discussion and Introduction and included two fabricated claims (a GRADE-pipeline

mandate mis-attributed to Di Pumpo et al. [87] and Forero et al. [35], and a misattributed processing-time figure for Tripathi et al. [5]), reattribution of a corpus-doubling claim from Torres et al. [2] to Trad et al. [1], and a reversed pair of κ values attributed to Zuo et al. [78].

Appendix B.5. Notable Corrections

Supplementary Table S8 documents the twelve most significant corrections (those involving fabricated claims, numerical transpositions, or wrong-paper attributions).

Appendix B.6. Resolution Status

Every key-slot reached one of three human-recorded resolution states at audit close: **Verified** (541 slots, 88.4%)—the notebook query confirmed the claim as stated, no edit required; **Corrected** (36 slots, 5.9%)—the manuscript was edited to replace an inaccurate or mis-attributed claim with a source-grounded sentence, then re-audited; and **Accepted** (35 slots, 5.7%)—partially accurate claims accepted after minor adjustment, together with the complementary references validated against their own sources. No key-slot remained unresolved.

The complete row-by-row audit log is provided in the accompanying `Citation_Audit_Log.xlsx` workbook (Sheet: Audit Log), deposited at Zenodo, which records the verbatim claim text, verdict, issue identified, recommended correction, verification source, and resolution status for every key-slot.

Appendix C. Cross-Cutting Evidence Map

Table A1 provides the full cross-cutting evidence map introduced in Section 8 and summarised visually in Figure 14: challenge dimension × stage(s) affected × best current approach × remaining gap. Stages A–G correspond to the PRISMA taxonomy used throughout: A = search, B = screening, C = full-text selection, D = quality appraisal, E = extraction, F = synthesis and reporting, G = agentic pipelines.

Table A1. Cross-cutting synthesis matrix.

Challenge	Stage(s)	Best Current Approach (Corpus)	Remaining Gap
Calibration and confidence misestimation	All (most severe in B, D)	Likert/numeric threshold scoring via Youden's index [29]; one-shot inclusivity-adjustment threshold modulation [61]; three-way Include/Exclude/Uncertain classification [77]; compact LLM scoring at 50–75 rating thresholds [69]	ROBIS agreement near-chance: $\kappa = 0.297$ for best model across 20 SRs [35]; PRISMA 2020 adherence estimates differed from human experts by 23.1–29.7%; PRIOR adherence estimates showed smaller differences of 6.1–14.0% [105]; no domain-general calibration method demonstrated
Hallucination and value fabrication	B, E, F (citations in F)	Retrieval-Augmented Generation with uploaded corpus [1]; intermediate "study summary" extraction step to filter noise before data tables [90]; parameter calibration via low temperature and top-p sampling settings [114]	Citation hallucination rates: GPT-3.5 39.6%, GPT-4 28.6%, Bard 91.4% [108]; of 22 studies retrieved by Consensus/GPT-4, 1 was a complete hallucination, 8 had author errors, 7 had publication-year errors [111]; LLMs can generate fabricated or hallucinated data items absent from source texts [93]; no automatic hallucination detection with acceptable recall demonstrated
Human–AI inter-rater agreement	B, C, D	Multi-model parallel extraction reconciled by a human expert [128]; dual-run intra-AI verification prompting to reduce false negatives [117]; zero-shot prompting with explicitly listed exclusion criteria [71]; proposed human-in-the-loop parallel human-and-AI review of every record [52]	Agreement varies widely across tasks: $\kappa = 0.895$ for data extraction in second-language acquisition research [53] vs $\kappa = 0.186$ for systematic review risk-of-bias assessment using ROBIS [35]; full-text and RoB instruments consistently below $\kappa = 0.60$; no agreed acceptable-κ threshold for AI-assisted reviewing

Table A1. *Cont.*

Challenge	Stage(s)	Best Current Approach (Corpus)	Remaining Gap
Reproducibility and versioning	All	Disclosure of the exact model version and deployment platform utilised [5]; verbatim prompt release [49]; temperature = 1 and fixed random seed to characterise non-determinism	Query recall variance 12.0% across 10 identical runs vs 7.1% for precision [42]; rapid model evolution means identical prompts may yield different results in future versions [41]; temperature undisclosed in the great majority of the corpus; proprietary model opacity may affect reproducibility [9]
Domain and language generalisation	A, B, E	Domain-specific fine-tuning [37]; domain-vocabulary prompt augmentation	Model-version effects on MeSH term coverage documented across prompting rounds [44]; performance gap widens in non-English and specialist domains; no validated cross-domain benchmark for LLM-assisted reviewing
Context-window management	C, E, G (full-text)	Retrieval-Augmented Generation avoiding text segmentation to preserve document coherence [1]	Session-level batching (5 abstracts/3 full texts) acknowledged to "hinder the model's ability to grasp broader document context" [94]; strict token limits impaired risk-of-bias judgements [88]; prompt-length constraints inherently restrict LLM context associations [4]; no standard chunking or RAG protocol
Error propagation in pipelines	G (agentic)	Breaking down the review into manageable steps to incorporate human validation at each step [13]; no study ever included or excluded solely on automated classification [54]; comprehensive content validation distinguishing fact-checking from full content validation [96]	T/A sensitivity of 54.9% propagates to exclude nearly half of relevant literature from full-text and meta-analysis stages [11]; PROMPTHEUS achieved ROUGE-1 precision of 0.969–0.995 but recall of only 0.026–0.039 [2]; downstream data extraction inaccuracies—such as omission of data from a critical outcome—could significantly sway a review's conclusions [93]; no automated inter-stage error-detection mechanism demonstrated in any corpus pipeline
Reporting transparency and reproducibility	F (and all stages for PRISMA 2020)	RAISE and DEST disclosure guidelines for Generative AI manuscripts [22]; CONSORT-AI and PRISMA-AI aligned reporting with exact LLM version, access mode, context-window length, and full prompt history [26]; explicit parameter reporting (temperature, top-*p*) [5]	Heavy reliance on proprietary models with opaque decision making and frequent updates may affect reproducibility [9]; temperature undisclosed in the vast majority of the corpus; no versioned prompt-registry standard adopted

References

1. Trad, F.; Yammine, R.; Charafeddine, J.; Chakhtoura, M.; Rahme, M.; El-Hajj Fuleihan, G.; Chehab, A. Streamlining systematic reviews with large language models using prompt engineering and retrieval augmented generation. *BMC Med. Res. Methodol.* **2025**, *25*, 130. [CrossRef] [PubMed]
2. Torres, J.; Mulligan, C.; Jorge, J.; Moreira, C. PROMPTHEUS: A Human-Centered Pipeline to Streamline Systematic Literature Reviews with Large Language Models. *Information* **2025**, *16*, 420. [CrossRef]
3. Hampson, T.; Cargos, K.; McKinley, J. Automate the 'boring bits': An assessment of AI-assisted systematic review (AIASR). *Res. Methods Appl. Linguist.* **2025**, *4*, 100258. [CrossRef]
4. Luo, X.; Chen, F.; Zhu, D.; Wang, L.; Wang, Z.; Liu, H.; Lyu, M.; Wang, Y.; Wang, Q.; Chen, Y. Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses. *J. Med. Internet Res.* **2024**, *26*, e56780. [CrossRef] [PubMed]
5. Tripathi, M.; Liapis, I.; Sanghera, J.; Annesi, C.; Landrum, A.; Kumar, S.; Bhatia, S.; Fonseca, A. Leveraging large language models to automate scoping reviews: A case study in treatment options for pancreatic cancer. *Am. J. Surg.* **2026**, *252*, 116764. [CrossRef] [PubMed]
6. Erslev, M.; Tretow-Fish, T. From Bullshit to Cognition: Computing Within the Epistemic Crisis of Large Language Models in Systematic Literature Review. In Proceedings of the Conference Proceedings—Computing X Crisis: 6th Decennial Aarhus Conference, AAR 2025, Aarhus, Denmark, 18–22 August 2025; pp. 272–280. [CrossRef]
7. Köhler, J.; Harl, M.V.; Westner, M.; Strahringer, S. Can AI be a Scholar? A Systematic Review on the Role of Generative AI in Systematic Literature Reviews. In Proceedings of the 2025 27th International Conference on Business Informatics (CBI), Lisbon, Portugal, 9–12 September 2025; pp. 1–11. [CrossRef]
8. Scherbakov, D.; Hubig, N.; Jansari, V.; Bakumenko, A.; Lenert, L.A. The emergence of large language models as tools in literature reviews: A large language model-assisted systematic review. *J. Am. Med. Inform. Assoc. JAMIA* **2025**, *32*, 1071–1086. [CrossRef] [PubMed]

9. Lieberum, J.L.; Töws, M.; Metzendorf, M.I.; Heilmeyer, F.; Siemens, W.; Haverkamp, C.; Böhringer, D.; Meerpohl, J.; Eisele-Metzger, A. Large language models for conducting systematic reviews: On the rise, but not yet ready for use—A scoping review. *J. Clin. Epidemiol.* **2025**, *181*, 111746. [CrossRef] [PubMed]
10. Sollini, M.; Pini, C.; Lazar, A.; Gelardi, F.; Ninatti, G.; Bauckneht, M.; Chiti, A.; Kirienko, M. Human researchers are superior to large language models in writing a medical systematic review in a comparative multitask assessment. *Sci. Rep.* **2026**, *16*, 173. [CrossRef] [PubMed]
11. Purewal, A.; Fautsch, K.; Klasova, J.; Hussain, N.; D'Souza, R. Human versus artificial intelligence: Evaluating ChatGPT's performance in conducting published systematic reviews with meta-analysis in chronic pain research. *Reg. Anesth. Pain Med.* **2025**, *51*, 437–442. [CrossRef] [PubMed]
12. Roy, K.; Khandelwal, V.; Vera, V.; Surana, H.; Heckman, H.; Sheth, A. GEAR-Up: Generative AI and External Knowledge-based Retrieval Upgrading Scholarly Article Searches for Systematic Reviews. *Proc. Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 23823–23825. [CrossRef]
13. Bayani, A.; Epoh Ewane, L.; Oliveira dos Anjos, D.; Mac-Seing, M.; Nikiema, J. Leveraging open-source large language models (LLMs) in scoping reviews: A case study on disability and AI applications. *Int. J. Med. Inform.* **2025**, *204*, 106048. [CrossRef] [PubMed]
14. Felizardo, K.; Lima, M.; Deizepe, A.; Conte, T.; Steinmacher, I. ChatGPT application in Systematic Literature Reviews in Software Engineering: An evaluation of its accuracy to support the selection activity. In Proceedings of the International Symposium on Empirical Software Engineering and Measurement, Barcelona, Spain, 24–25 October 2024; pp. 25–36. [CrossRef]
15. Kelley, C.; Kelley, S.; Rivas-Koehl, M.; Wilson, A.; Rosinsky, K.; Briggs, S.; Scott, M.E. Identifying and Prioritizing Core Components of Relationship Education Programs: A Case Study of an Artificial Intelligence (AI) Assisted Systematic Review. *Prev. Sci.* **2026**, *27*, 1191–1202. [CrossRef] [PubMed]
16. Schopow, N.; Osterhoff, G.; Baur, D. Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review. *JMIR Med. Inform.* **2023**, *11*, e48933. [CrossRef] [PubMed]
17. Kim, J.; Rickard, M.; Dangle, P.; Batra, N.; Chua, M.; Khondker, A.; Szymanski, K.; Misseri, R.; Lorenzo, A. Evaluating large language models for title/abstract screening: A systematic review and meta-analysis & development of new tool. *J. Med. Artif. Intell.* **2025**, *8*, 34. [CrossRef]
18. Delgado-Chaves, F.; Jennings, M.; Atalaia, A.; Wolff, J.; Horvath, R.; Mamdouh, Z.; Baumbach, J.; Baumbach, L. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proc. Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2411962122. [CrossRef] [PubMed]
19. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLoS Med.* **2021**, *18*, e1003583. [CrossRef] [PubMed]
20. Thomas, J.; Harden, A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med. Res. Methodol.* **2008**, *8*, 45. [CrossRef] [PubMed]
21. OpenAI. GPT-4 Technical Report. Version 6, updated 2024. *arXiv* **2024**, arXiv:2303.08774.
22. Farotimi, O.; Dunn, A.; Van Lissa, C.J.; Polanin, J.R.; Mavridis, D.; Pigott, T.D. Guidance for manuscript submissions testing the use of generative AI for systematic review and meta-analysis. *Res. Synth. Methods* **2026**, *17*, 237–239. [CrossRef] [PubMed]
23. Hong, Q.N.; Fàbregues, S.; Bartlett, G.; Boardman, F.; Cargo, M.; Dagenais, P.; Gagnon, M.P.; Griffiths, F.; Nicolau, B.; O'Cathain, A.; et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Educ. Inf.* **2018**, *34*, 285–291.. [CrossRef]
24. Yun, H.; Trikalinos, T.; Marshall, I.; Wallace, B. Appraising the Potential Uses and Harms of Large Language Models for Medical Systematic Reviews. In Proceedings of the EMNLP 2023—2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 10122–10139. [CrossRef]
25. Soprano, M.; Modha, S.; Roitero, K.; Maddalena, E.; Viviani, M.; Pasi, G.; Mizzaro, S. AIDME: A Scalable, Interpretable Framework for AI-Aided Scoping Reviews. In Proceedings of the ICTIR 2025—Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, Padua, Italy, 18 July 2025; pp. 194–207. [CrossRef]
26. Galli, C.; Gavrilova, A.; Calciolari, E. Large Language Models in Systematic Review Screening: Opportunities, Challenges, and Methodological Considerations. *Information* **2025**, *16*, 378. [CrossRef]
27. Homiar, A.; Thomas, J.; Ostinelli, E.G.; Kennett, J.; Friedrich, C.; Cuijpers, P.; Harrer, M.; Leucht, S.; Miguel, C.; Rodolico, A.; et al. Development and evaluation of prompts for a large language model to screen titles and abstracts in a living systematic review. *BMJ Ment. Health* **2025**, *28*, e301762. [CrossRef] [PubMed]
28. Insuk, S.; Boonpattharatthiti, K.; Booncharoen, C.; Chaipitak, P.; Rashid, M.; Veettil, S.K.; Lai, N.M.; Chaiyakunapruk, N.; Dhippayom, T. How Well Do ChatGPT and Claude Perform in Study Selection for Systematic Review in Obstetrics. *J. Med. Syst.* **2025**, *49*, 110. [CrossRef] [PubMed]

29. Issaiy, M.; Ghanaati, H.; Kolahi, S.; Shakiba, M.; Jalali, A.; Zarei, D.; Kazemian, S.; Avanaki, M.; Firouznia, K. Methodological insights into ChatGPT's screening performance in systematic reviews. *BMC Med. Res. Methodol.* **2024**, *24*, 78. [CrossRef] [PubMed]
30. Oami, T.; Okada, Y.; Nakada, T.A. GPT-3.5 Turbo and GPT-4 Turbo in Title and Abstract Screening for Systematic Reviews. *JMIR Med. Inform.* **2025**, *13*, e64682. [CrossRef] [PubMed]
31. Gartlehner, G.; Kahwati, L.; Hilscher, R.; Thomas, I.; Kugley, S.; Crotty, K.; Viswanathan, M.; Nussbaumer-Streit, B.; Booth, G.; Erskine, N.; et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Res. Synth. Methods* **2024**, *15*, 576–589. [CrossRef] [PubMed]
32. Motzfeldt Jensen, M.; Brix Danielsen, M.; Riis, J.; Assifuah Kristjansen, K.; Andersen, S.; Okubo, Y.; Grønbech Jørgensen, M. ChatGPT-4o can serve as the second rater for data extraction in systematic reviews. *PLoS ONE* **2025**, *20*, e0313401. [CrossRef] [PubMed]
33. Kasireddy, E.; Chow, C.; Collet, J.; Pourrahmat, M.M.; Fazeli, M. Evaluating the Performance of Claude 3.7 Sonnet in Data Extraction Automation for Systematic Literature Reviews. *Value Health Reg. Issues* **2026**, *53*, 101539. [CrossRef] [PubMed]
34. Romano, G.; Romano, E.; Rau, M. Can Large Language Models Replicate Systematic Review Outcome Classifications in Medical Education? A Pilot Study Using Kirkpatrick Levels. *Med. Sci. Educ.* **2026**, *36*, 11–15. [CrossRef] [PubMed]
35. Forero, D.; Abreu, S.; Tovar, B.; Oermann, M. Automated analyses of risk of bias and critical appraisal of systematic reviews (ROBIS and AMSTAR 2): A comparison of the performance of 4 large language models. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 1471–1476. [CrossRef] [PubMed]
36. Srinivasan, A.; Berkowitz, J.; Friedrich, N.A.; Kivelson, S.; Tatonetti, N.P. Large Language Model Analysis of Reporting Quality of Randomized Clinical Trial Articles: A Systematic Review. *JAMA Netw. Open* **2025**, *8*, e2529418. [CrossRef] [PubMed]
37. Adam, G.; Deyoung, J.; Paul, A.; Saldanha, I.; Balk, E.; Trikalinos, T.; Wallace, B. Literature search sandbox: A large language model that generates search queries for systematic reviews. *JAMIA Open* **2024**, *7*, ooae098. [CrossRef] [PubMed]
38. Bourgeois, J.P.; Ellingson, H. Ability of ChatGPT to Generate Systematic Review Search Strategies Compared to a Published Search Strategy. *Med. Ref. Serv. Q.* **2025**, *44*, 279–291. [CrossRef] [PubMed]
39. Gitman, V.; Maxwell, C.; Gamble, J.M. Enhancing search strategies for systematic reviews on drug Harms: An evaluation of the utility of ChatGPT in error detection and keyword generation. *Comput. Biol. Med.* **2025**, *193*, 110464. [CrossRef] [PubMed]
40. Park, S.; Carroll, M.; Esteve, L.; Singh, K. Exploring Generative AI and Natural Language Processing to Develop Search Strategies for Systematic Reviews. In Proceedings of the ASEE Annual Conference and Exposition, Conference Proceedings, Portland, OR, USA, 23–26 June 2024. [CrossRef]
41. Tam, W.S.W.; Tung, N.; Lee, S.X.; Štiglic, G.; Huynh, T.; Tang, A. Development and Evaluation of a Generative AI Chatbot for Database Searching in Systematic Review. *J. Nurs. Sch.* **2026**, *58*, e70076. [CrossRef] [PubMed]
42. Wang, S.; Scells, H.; Koopman, B.; Zuccon, G. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? In Proceedings of the SIGIR 2023—Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, 23–27 July 2023; pp. 1426–1436. [CrossRef]
43. Wang, Z.; Cao, L.; Danek, B.; Jin, Q.; Lu, Z.; Sun, J. Accelerating clinical evidence synthesis with large language models. *npj Digit. Med.* **2025**, *8*, 509. [CrossRef] [PubMed]
44. Yu, F.; Kincaide, H.; Carlson, R. An Empirical Study Evaluating ChatGPT's Performance in Generating Search Strategies for Systematic Reviews. *Proc. Proc. Assoc. Inf. Sci. Technol.* **2024**, *61*, 423–434. [CrossRef]
45. Balel, Y.; Zogo, A.; Yıldız, S.; Tanyeri, H. Can ChatGPT-4o provide new systematic review ideas to oral and maxillofacial surgeons? *J. Stomatol. Oral Maxillofac. Surg.* **2024**, *125*, 101979. [CrossRef] [PubMed]
46. Ruksakulpiwat, S.; Phianhasin, L.; Benjasirisan, C.; Ding, K.; Ajibade, A.; Kumar, A.; Stewart, C. Assessing the Efficacy of ChatGPT Versus Human Researchers in Identifying Relevant Studies on mHealth Interventions for Improving Medication Adherence in Patients with Ischemic Stroke When Conducting Systematic Reviews: Comparative Analysis. *JMIR mHealth uHealth* **2024**, *12*, e51526. [CrossRef] [PubMed]
47. Bakare, A.; Lee, Y.; Hong, J.; Richter, C.P.; Kuriakose, J. Assessing Large Language Models for Early Article Identification in Otolaryngology—Head and Neck Surgery Systematic Reviews. *Health Care Sci.* **2026**, *5*, 19–28. [CrossRef] [PubMed]
48. Chen, H.; Jiang, Z.; Liu, X.; Xue, C.; Yew, S.; Sheng, B.; Zheng, Y.F.; Wang, X.; Wu, Y.; Sivaprasad, S.; et al. Can large language models fully automate or partially assist paper selection in systematic reviews? *Br. J. Ophthalmol.* **2025**, *109*, 962–966. [CrossRef] [PubMed]
49. Masinde, M. Enhancing Systematic Literature Reviews using LDA and ChatGPT: Case of Framework for Smart City Planning. In Proceedings of the 2024 IST-Africa Conference, IST-Africa 2024, Dublin, Ireland, 20–24 May 2024. [CrossRef]
50. Liang, L.; Wang, X.; Zhang, X.; Jia, X.; Chen, Z.; Zheng, X. A systematic review of photovoltaic power plant site selection approaches: Literature retrieval and analysis method based on ChatGPT and DeepSeek. *Energy Rep.* **2025**, *14*, 1993–2014. [CrossRef]
51. Akinseloyin, O.; Jiang, X.; Palade, V. Large language model-based multiagent collaboration for abstract screening toward automated systematic reviews. *Biol. Methods Protoc.* **2026**, *11*, bpag006. [CrossRef] [PubMed]

52. Addison, S.; Campion, E.; Landsman, M.; Rogalin, C.; Veeh, C. Making AI Accessible to Social Work Researchers: An Exploratory Analysis of Using ChatGPT to Screen Articles for Systematic and Scoping Reviews. *Res. Soc. Work Pract.* **2025**, *36*, 364–381. [CrossRef]
53. Aryadoust, V.; Jia, Y. Univariate normality checking practices in L2 research: An AI-assisted systematic review. *Stud. Second Lang. Acquis.* **2026**, *48*, 535–570. [CrossRef]
54. Borges, F.T.; Machado, G.D.M.; Santana, M.A.D.; Sancho, K.A.; França, G.V.A.D.; Santos, W.P.D.; Siqueira, C.E.G. LLM-Assisted Scoping Review of Artificial Intelligence in Brazilian Public Health: Lessons from Transfer and Federated Learning for Resource-Constrained Settings. *Int. J. Environ. Res. Public Health* **2026**, *23*, 81. [CrossRef] [PubMed]
55. Castillo-Segura, P.; Alario-Hoyos, C.; Kloos, C.; Fernandez Panadero, C. Leveraging the Potential of Generative AI to Accelerate Systematic Literature Reviews: An Example in the Area of Educational Technology. In Proceedings of the 2023 IEEE IFEES World Engineering Education Forum and Global Engineering Deans Council: Convergence for a Better World: A Call to Action, WEEF-GEDC 2023—Proceedings, Monterrey, Mexico, 23–27 October 2023. [CrossRef]
56. Chelstrom, B.; Polak, M.; Morgan, D.; Henak, C. Mechanical properties of the meninges: An expansive LLM-assisted systematic review. *Mater. Today Commun.* **2025**, *49*, 114034. [CrossRef]
57. Herbst, P.; Baars, H. Accelerating literature screening for systematic literature reviews with Large Language Models—Development, application, and first evaluation of a solution. In Proceedings of the CEUR Workshop Proceedings, Marburg, Germany, 9–11 October 2023.
58. Ito, Y.; Ikehara, H.; Okamoto, Y.; Kako, J. Is Large Language Model-Assisted Citation Screening Feasible in a Scoping Review on Nonpharmacological Interventions for Delirium in Patients With Cancer? *Cureus* **2025**, *17*, e90026. [CrossRef] [PubMed]
59. Jaumann, C.; Wiedholz, A.; Friedrich, A. LGAR: Zero-Shot LLM-Guided Neural Ranking for Abstract Screening in Systematic Literature Reviews. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 27 July–1 August 2025; pp. 7910–7927. [CrossRef]
60. Kadir, S.A.; Baranidharan, G.; Jiffry, A.; Krishnaswamy, G.; Sindrilla, M. AI-Assisted Systematic Review on the Impact of Protein Supplementation for Muscle Building on Kidney Function in Healthy Adults. In Proceedings of the 2025 International Conference on Recent Innovation in Science Engineering and Technology (ICRISET), Chennai, India, 1–2 August 2025; pp. 1–8. [CrossRef]
61. Kim, S.; Yoon, H.J. Large Language Model-Assisted Systematic Review: Validation Based on Cochrane Review Data. *Stud. Health Technol. Inform.* **2025**, *327*, 904–905. [CrossRef] [PubMed]
62. López-Pineda, A.; Nouni-García, R.; Carbonell-Soliva, Á.; Gil-Guillén, V.; Carratalá-Munuera, C.; Borrás, F. Validation of large language models (Llama 3 and ChatGPT-4o mini) for title and abstract screening in biomedical systematic reviews. *Res. Synth. Methods* **2025**, *16*, 620–630. [CrossRef] [PubMed]
63. Manullang, A.; Gao, X.; Viavattene, C.; Rahmanti, A. Comparative Assessment of ChatGPT, DeepSeek, and Human Reviewers for Full-Text Screening in Systematic Reviews on the Impact of Air Pollution in Respiratory Diseases. In *Artificial Intelligence XLII*; Bramer, M., Stahl, F., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2026; Volume 16302, pp. 478–484. [CrossRef]
64. Matsui, K.; Utsumi, T.; Aoki, Y.; Maruki, T.; Takeshima, M.; Takaesu, Y. Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews. *J. Med. Internet Res.* **2024**, *26*, e52758. [CrossRef] [PubMed]
65. Nordmann, K.; Schaller, M.; Sauter, S.; Fischer, F. Capability of chatbots powered by large language models to support the screening process of scoping reviews: A feasibility study. *JAMIA Open* **2026**, *9*, ooaf098. [CrossRef] [PubMed]
66. Nordmann, K.; Fischer, F. Harnessing ChatGPT for abstract screening in health-related scoping reviews: The role of structured eligibility criteria. *BMC Health Serv. Res.* **2026**, *26*, 109. [CrossRef] [PubMed]
67. Oami, T.; Okada, Y.; Nakada, T.A. Optimal large language models to screen citations for systematic reviews. *Res. Synth. Methods* **2025**, *16*, 859–875. [CrossRef] [PubMed]
68. Rokhshad, R.; Bagherianlemraski, M.; Ehsani, S.; Haghighat, S.; Schwendicke, F. Large language models for the screening step in systematic reviews in dentistry. *J. Dent.* **2025**, *160*, 105877. [CrossRef] [PubMed]
69. Sciurti, A.; Migliara, G.; Siena, L.; Isonne, C.; De Blasiis, M.; Sinopoli, A.; Iera, J.; Marzuillo, C.; De Vito, C.; Villari, P.; et al. Compact large language models for title and abstract screening in systematic reviews: An assessment of feasibility, accuracy, and workload reduction. *Res. Synth. Methods* **2026**, *17*, 332–347. [CrossRef] [PubMed]
70. Sung, H.; Altahsh, D.; Garrison, S. AI-Assisted Systematic Review: Humans Still Need to Review All Abstracts for Inclusion. *JMIR Form. Res.* **2026**, *10*, e82896. [CrossRef] [PubMed]
71. Syriani, E.; David, I.; Kumar, G. Screening articles for systematic reviews with ChatGPT. *J. Comput. Lang.* **2024**, *80*, 101287. [CrossRef]
72. Vallamchetla, S.K.; Abdelkader, O.; Elnaggar, A.; Ramadan, D.; Islam Shourav, M.M.; Riaz, I.B.; Lin, M.P. Do it faster with PICOS: Generative AI-Assisted systematic review screening. *J. Biomed. Inform.* **2025**, *168*, 104860. [CrossRef] [PubMed]

73. van Mossel, S.; Oude-Wolcherink, M.; de Feria Cardet, R.; de Geus-Oei, L.F.; Vriens, D.; Koffijberg, H.; Saing, S. Artificial Intelligence as a New Research Ally? Performing AI-Assisted Systematic Literature Reviews in Health Economics. *PharmacoEconomics* **2025**, *43*, 647–650. [CrossRef] [PubMed]
74. Valizadeh, A.; Moassefi, M.; Nakhostin-Ansari, A.; Hosseini Asl, S.; Saghab Torbati, M.; Aghajani, R.; Maleki Ghorbani, Z.; Faghani, S. Abstract screening using the automated tool Rayyan: Results of effectiveness in three diagnostic test accuracy systematic reviews. *BMC Med. Res. Methodol.* **2022**, *22*, 160. [CrossRef] [PubMed]
75. Wang, S.; Scells, H.; Zhuang, S.; Potthast, M.; Koopman, B.; Zuccon, G. Zero-Shot Generative Large Language Models for Systematic Review Screening Automation. In *Advances in Information Retrieval*; Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2024; Volume 14608 LNCS, pp. 403–420. [CrossRef]
76. Yu, F.; Liu, C.; Sharmin, S. Performance, Usability, and User Experience of Rayyan for Systematic Reviews. *Proc. Proc. Assoc. Inf. Sci. Technol.* **2022**, *59*, 843–844. [CrossRef]
77. Zrubka, M.; Alexy, M.; Gyorgy, K.; Zrubka, Z. Large Language Models for Title/Abstract Screening in Systematic Literature Reviews: A Case Study in Precision Livestock Farming. In Proceedings of the CINTI 2025—IEEE 25th International Symposium on Computational Intelligence and Informatics, Proceedings, Budapest, Hungary, 18–20 November 2025; pp. 661–666. [CrossRef]
78. Zuo, C.; Yang, X.; Errickson, J.; Li, J.; Hong, Y.; Wang, R. AI-assisted evidence screening method for systematic reviews in environmental research: Integrating ChatGPT with domain knowledge. *Environ. Evid.* **2025**, *14*, 5. [CrossRef] [PubMed]
79. Cao, C.; Sang, J.; Arora, R.; Chen, D.; Kloosterman, R.; Cecere, M.; Gorla, J.; Saleh, R.; Drennan, I.; Teja, B.; et al. Development of Prompt Templates for Large Language Model–Driven Screening in Systematic Reviews. *Ann. Intern. Med.* **2025**, *178*, 389–401. [CrossRef] [PubMed]
80. Dogra, S.; Arabshahi, S.; Wei, J.; Hu, E.; Saidenberg, L.; Sharma, S.; Gu, Z.; Siriruchatanon, M.; Kang, S. Evaluating Large Language Models for Radiology Systematic Review Title and Abstract Screening. *Acad. Radiol.* **2025**, *32*, 7023–7034. [CrossRef] [PubMed]
81. Doneva, S.; De Viragh, S.; Hubarava, H.; Schandelmaier, S.; Briel, M.; Ineichen, B. StudyTypeTeller—Large language models to automatically classify research study types for systematic reviews. *Res. Synth. Methods* **2025**, *16*, 1005–1024. [CrossRef] [PubMed]
82. Ramchandani, R.; Guo, E.; Rakab, E.; Rathod, J.; Strain, J.; Klement, W.; Shorr, R.; Williams, E.; Jones, D.; Gilbert, S. Validation of automated paper screening for esophagectomy systematic review using large language models. *PeerJ Comput. Sci.* **2025**, *11*, e2822. [CrossRef] [PubMed]
83. Tran, V.T.; Gartlehner, G.; Yaacoub, S.; Boutron, I.; Schwingshackl, L.; Stadelmaier, J.; Sommer, I.; Alebouyeh, F.; Afach, S.; Meerpohl, J.; et al. Sensitivity and Specificity of Using GPT-3.5 Turbo Models for Title and Abstract Screening in Systematic Reviews and Meta-analyses. *Ann. Intern. Med.* **2024**, *177*, 791–799. [CrossRef] [PubMed]
84. Tran, V.T.; Grana Possamai, C.; Boutron, I.; Ravaud, P. Using large language models to directly screen electronic databases as an alternative to traditional search strategies such as the Cochrane highly sensitive search for filtering randomized controlled trials in systematic reviews. *Res. Synth. Methods* **2025**, *16*, 1035–1041. [CrossRef] [PubMed]
85. Vembye, M.; Christensen, J.; Mølgaard, A.; Schytt, F. Generative Pretrained Transformer Models Can Function as Highly Reliable Second Screeners of Titles and Abstracts in Systematic Reviews: A Proof of Concept and Common Guidelines. *Psychol. Methods* 2025, *online ahead of print*. [CrossRef] [PubMed]
86. Chen, S.F.; Alyakin, A.; Seas, A.; Yang, E.; Choi, J.J.; Lee, J.V.; Chen, A.L.; Warman, P.I.; Bitolas, R.T.; Steele, R.J.; et al. LLM-assisted systematic review of large language models in clinical medicine. *Nat. Med.* **2026**, *32*, 1152–1159. [CrossRef] [PubMed]
87. Di Pumpo, M.; Riccardi, M.T.; De Vita, V.; Damiani, G. Evaluation of a large language model (ChatGPT) versus human researchers in assessing risk-of-bias and community engagement levels: A systematic review use-case analysis. *Eur. J. Public Health* **2025**, *35*, 1082–1086. [CrossRef] [PubMed]
88. Hasan, B.; Saadi, S.; Rajjoub, N.; Hegazi, M.; Al-Kordi, M.; Fleti, F.; Farah, M.; Riaz, I.; Banerjee, I.; Wang, Z.; 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*, 394–398. [CrossRef] [PubMed]
89. Kim, J.K.; Chua, M.E.; Li, T.G.; Rickard, M.; Lorenzo, A.J. Novel AI applications in systematic review: GPT-4 assisted data extraction, analysis, review of bias. *BMJ Evid.-Based Med.* **2025**, *30*, 313–322. [CrossRef] [PubMed]
90. Teperikidis, L.; Boulmpou, A.; Papadopoulos, C.; Biondi-Zoccai, G. Using ChatGPT to perform a systematic review: A tutorial. *Minerva Cardiol. Angiol.* **2024**, *72*, 547–567. [CrossRef] [PubMed]
91. Yang, R.; Tong, J.; Wang, H.; Huang, H.; Hu, Z.; Li, P.; Liu, N.; Lindsell, C.; Pencina, M.; Chen, Y.; et al. Enabling inclusive systematic reviews: Incorporating preprint articles with large language model-driven evaluations. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 1718–1725. [CrossRef] [PubMed]
92. Yao, B.; Ergun, O.; Ding, M.; Miller, C.; Narayan, V.; Dahm, P. Assessing the methodologic quality of systematic reviews using generative large language models. *Can. Urol. Assoc. J.* **2025**, *19*, E427–E433. [CrossRef] [PubMed]

93. Gartlehner, G.; Kahwati, L.; Nussbaumer-Streit, B.; Crotty, K.; Hilscher, R.; Kugley, S.; Viswanathan, M.; Thomas, I.; Konet, A.; Booth, G.; et al. From promise to practice: Challenges and pitfalls in the evaluation of large language models for data extraction in evidence synthesis. *BMJ Evid.-Based Med.* **2025**, *30*, 385–389. [CrossRef] [PubMed]
94. 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*, 616–626. [CrossRef] [PubMed]
95. Peer, J.; Vadivel, V.S. A New Systematic Literature Review Approach using Rayyan and Copilot on Apps used by persons with special needs. In Proceedings of the 2025 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE), Singapore, 4–7 November 2025; pp. 1–7. [CrossRef]
96. Pham, V.; Lin, F.R. The Design and Evaluation of the Collaboration between Researchers and Generative AI for Systematic Literature Reviews. In Proceedings of the Annual Hawaii International Conference on System Sciences, Waikoloa Village, HI, USA, 7–10 January 2025; pp. 4685–4694. [CrossRef]
97. Sampaio, R.; Chagas, V.; Sanchez, C.; Gonçalves, J.; Borges, T.; Alison, M.; Tigrinho, C.; de Souza, J.; Paz, F. An Artificial Intelligence (AI)-Assisted Scoping Review of Emerging Uses of AI in Qualitative Research and Its Ethical Considerations. *Rev. Pesqui. Qual.* **2024**, *12*, 1–26. [CrossRef]
98. Shah, A.; Mehendale, S.; Kanthi, S. Efficacy of Large Language Models for Systematic Reviews. In Proceedings of the 2024 2nd International Conference on Foundation and Large Language Models, FLLM 2024, Dubai, United Arab Emirates, 26–29 November 2024; pp. 29–35. [CrossRef]
99. Kaur, R.; Singh, S. A large language model framework for knowledge graph construction of randomized controlled trials for evidence synthesis and querying. *IISE Trans. Healthc. Syst. Eng.* **2026**, *16*, 131–147. [CrossRef]
100. Tomassini, F.; Luraschi, A.; Patarnello, S.; Masciocchi, C.; Arcuri, G.; Lilli, L. Leveraging ChatGPT-4 for Evidence Synthesis: A Case Study on the Use of a Large Language Model in a Systematic Review. *Stud. Health Technol. Inform.* **2025**, *332*, 22–26. [CrossRef] [PubMed]
101. Holm, S.; Huss, F.; Nayyer, B.; Zdolsek, J. Use of Artificial Intelligence in Burn Assessment: A Scoping Review with a Large Language Model-Generated Decision Tree. *Eur. Burn J.* **2026**, *7*, 4. [CrossRef] [PubMed]
102. Liu, H.; Peng, Y.; Weng, C. How Good Is ChatGPT for Medication Evidence Synthesis? *Stud. Health Technol. Inform.* **2023**, *302*, 1062–1066. [CrossRef] [PubMed]
103. Karakikes, A.; Kotis, K. AI-Assisted OSINT/SOCMINT for Safeguarding Borders: A Systematic Review. *Information* **2025**, *16*, 1095. [CrossRef]
104. Pavičić, J.; Marušić, A.; Buljan, I. Using ChatGPT to Improve the Presentation of Plain Language Summaries of Cochrane Systematic Reviews About Oncology Interventions: Cross-Sectional Study. *JMIR Cancer* **2025**, *11*, e63347. [CrossRef] [PubMed]
105. Forero, D.; Abreu, S.; Tovar, B.; Oermann, M. Large Language Models and the Analyses of Adherence to Reporting Guidelines in Systematic Reviews and Overviews of Reviews (PRISMA 2020 and PRIOR). *J. Med. Syst.* **2025**, *49*, 80. [CrossRef] [PubMed]
106. Alshami, A.; Elsayed, M.; Ali, E.; Eltoukhy, A.; Zayed, T. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems* **2023**, *11*, 351. [CrossRef]
107. Silva, N.; Wickramaarachchi, D. Enhancing Systematic Literature Reviews: Evaluating the Performance of LLM-Based Tools Across Key Systematic Literature Review Stages. In Proceedings of the 2025 5th International Conference on Advanced Research in Computing (ICARC), Belihuloya, Sri Lanka, 19–20 February 2025; pp. 1–6. [CrossRef]
108. Chelli, M.; Descamps, J.; Lavoué, V.; Trojani, C.; Azar, M.; Deckert, M.; Raynier, J.L.; Clowez, G.; Boileau, P.; Ruetsch-Chelli, C. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *J. Med. Internet Res.* **2024**, *26*, e53164. [CrossRef] [PubMed]
109. Kusa, W.; Scells, H.; Staudinger, M.; Hanbury, A. Leveraging Cochrane Systematic Literature Reviews for Prospective Evaluation of Large Language Models. In Proceedings of the CEUR Workshop Proceedings, Glasgow, UK, 28 March 2024.
110. Najafali, D.; Camacho, J.M.; Reiche, E.; Galbraith, L.G.; Morrison, S.D.; Dorafshar, A.H. Truth or Lies? The Pitfalls and Limitations of ChatGPT in Systematic Review Creation. *Aesthetic Surg. J.* **2023**, *43*, N654–N655. [CrossRef] [PubMed]
111. Tosi, D. Comparing Generative AI Literature Reviews Versus Human-Led Systematic Literature Reviews: A Case Study on Big Data Research. *Proc. IEEE Access* **2025**, *13*, 56210–56219. [CrossRef]
112. Wang, S.; Scells, H.; Koopman, B.; Zuccon, G. Reassessing Large Language Model Boolean Query Generation for Systematic Reviews. In Proceedings of the SIGIR 2025—Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, Padua, Italy, 13–18 July 2025; pp. 3296–3305. [CrossRef]
113. Bernard, N.; Sagawa, Y.; Bier, N.; Lihoreau, T.; Pazart, L.; Tannou, T. Using artificial intelligence for systematic review: The example of elicit. *BMC Med. Res. Methodol.* **2025**, *25*, 75. [CrossRef] [PubMed]
114. Demir, G.; Süküt, Y.; Duran, G.; Topsakal, K.; Görgülü, S. Enhancing systematic reviews in orthodontics: A comparative examination of GPT-3.5 and GPT-4 for generating PICO-based queries with tailored prompts and configurations. *Eur. J. Orthod.* **2024**, *46*, cjae011. [CrossRef] [PubMed]

115. Ashkanani, Z.; Mohtar, R.; Al-Enezi, S.; Smith, P.K.; Calabrese, S.; Ma, X.; Abdullah, M. AI-assisted systematic review on remediation of contaminated soils with PAHs and heavy metals. *J. Hazard. Mater.* **2024**, *468*, 133813. [CrossRef] [PubMed]
116. Noroozi, M.; Moghaddam, H.R.; Shah, A.; Charkhgard, H.; Sarkar, S.; Das, T.K.; Pohland, T. An AI-Assisted Systematic Literature Review of the Impact of Vehicle Automation on Energy Consumption. *Proc. IEEE Trans. Intell. Veh.* **2023**, *8*, 3572–3592. [CrossRef]
117. Schmitt, V. Disentangling patent quality: Using a large language model for a systematic literature review. *Scientometrics* **2025**, *130*, 267–311. [CrossRef]
118. Ndukwe, H.; Otukpa, E. Protocol for a scoping review examining the application of large language models in healthcare education and public health learning spaces. *PLoS ONE* **2026**, *21*, e0339594. [CrossRef] [PubMed]
119. Jung, Y.; Jin, S.H. Questioning the role of AI as collaborator: A systematic literature review of generative AI-supported knowledge construction. *Interact. Learn. Environ.* **2025**, *34*, 2901–2920. [CrossRef]
120. Robleto, E.; Shehadeh, L. Accelerating systematic reviews: A novel 1-wk screening protocol using rule-based automation with AI-assisted Python coding. *Am. J. Physiol. Heart Circ. Physiol.* **2025**, *329*, H1391–H1413. [CrossRef] [PubMed]
121. Giunti, G.; Doherty, C.P. Cocreating an Automated mHealth Apps Systematic Review Process With Generative AI: Design Science Research Approach. *JMIR Med. Educ.* **2024**, *10*, e48949. [CrossRef] [PubMed]
122. Konet, A.; Thomas, I.; Gartlehner, G.; Kahwati, L.; Hilscher, R.; Kugley, S.; Crotty, K.; Viswanathan, M.; Chew, R. Performance of two large language models for data extraction in evidence synthesis. *Res. Synth. Methods* **2024**, *15*, 818–824. [CrossRef] [PubMed]
123. Khan, M.; Ayub, U.; Ahmed Naqvi, S.; Rubab Khakwani, K.; Riaz Sipra, Z.; Raina, A.; Zhou, S.; He, H.; Saeidi, A.; Hasan, B.; et al. Collaborative large language models for automated data extraction in living systematic reviews. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 638–647. [CrossRef] [PubMed]
124. Landschaft, A.; Antweiler, D.; Mackay, S.; Kugler, S.; Rüping, S.; Wrobel, S.; Höres, T.; Allende-Cid, H. Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews. *Int. J. Med. Inform.* **2024**, *189*, 105531. [CrossRef] [PubMed]
125. Hilkenmeier, F.; Pelzer, M.; Stierle, C.; Fink-Lamotte, J. Evaluating the AI Tool "Elicit" as a Semi-Automated Second Reviewer for Data Extraction in Systematic Reviews: A Proof-of-Concept. *Soc. Sci. Comput. Rev.* 2025, *online ahead of print*. [CrossRef]
126. Liu, J.; Lai, H.; Zhao, W.; Huang, J.; Xia, D.; Liu, H.; Luo, X.; Wang, B.; Pan, B.; Hou, L.; et al. AI-driven evidence synthesis: Data extraction of randomized controlled trials with large language models. *Int. J. Surg.* **2025**, *111*, 2722–2726. [CrossRef] [PubMed]
127. da Costa Mineiro, A.; de Mello Valério, V.; da Silva Arantes, I.; Neves, S.; de Cassia Arantes, R. Helix innovation models: Systematic literature review with data analysis script by R software and ChatGPT. *Qual. Quant.* **2025**, *59*, 1353–1381. [CrossRef]
128. García-Torres, D.; Vicente Ripoll, M.A.; Fernández Peris, C.; Mira Solves, J.J. Enhancing Clinical Reasoning with Virtual Patients: A Hybrid Systematic Review Combining Human Reviewers and ChatGPT. *Healthcare* **2024**, *12*, 2241. [CrossRef] [PubMed]
129. Surles, S.; Noteboom, C. Information system cognitive bias classifications and fairness in machine learning: Systematic review using large language models. *Issues Inf. Syst.* **2024**, *25*, 491–506. [CrossRef]
130. Albalawi, F.; Khanagar, S.; Iyer, K.; Alhazmi, N.; Alayyash, A.; Alhazmi, A.; Awawdeh, M.; Singh, O. Evaluating the Performance of Artificial Intelligence-Based Large Language Models in Orthodontics—A Systematic Review and Meta-Analysis. *Appl. Sci.* **2025**, *15*, 893. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.