ARISMA: Guidelines for AI- and LLM-Assisted Systematic Reviews, Scoping Reviews, and Mapping Studies

Mahyar Tourchi Moghaddam[0000-0001-5028-7546] and Mina Alipour[0000-0002-6717-8976]

University of Southern Denmark, Odense 5230, DK
{mtmo,mial}@mmmi.sdu.dk

Abstract. Systematic reviews, scoping reviews, mapping studies, and related evidence syntheses are increasingly difficult to conduct with fully manual workflows as search volumes, update cycles, and synthesis requirements continue to expand. At the same time, artificial intelligence, machine learning, and large language models are rapidly entering review practice across query formulation, screening, extraction, categorization, appraisal support, and reporting. Yet the empirical evidence remains uneven, task-dependent, and insufficient to justify unconstrained automation. Existing standards such as PRISMA 2020, PRISMA-S, PRISMA-ScR, PRISMA-P, PRESS, and SWiM remain essential, but none provides an end-to-end operational standard for when AI use is methodologically appropriate, how it should be validated, which review decisions must remain human-led, and how AI involvement should be reported so that readers can audit it. This paper proposes ARISMA, an **A**I **R**eporting and **I**ntegration standard for **S**ystematic **M**ethods and **A**nalysis. ARISMA treats AI as an inspected, benchmarked, logged, and reversible assistant rather than an autonomous reviewer. It is built around one governing principle: every consequential scientific decision must remain human-interpretable, human-auditable, and human-accountable. The paper contributes a lifecycle taxonomy, process guidance, stepwise recommendations across the review pipeline, a governance and provenance model, a tool-support framework, an AI-integrated reporting checklist, and a validation matrix. It also addresses legal, privacy, infrastructure, and sustainability considerations. The framework was iteratively refined through structured expert consultation. The result is a practical and auditable guideline for responsible AI-assisted evidence synthesis.

Keywords: Evidence Synthesis · Systematic Review Methodology · LLM-assisted Review · Snowballing · Mapping Study · AI-assisted Review.

1 Introduction

Systematic reviews do not merely summarize literature; they define what counts as the current state of the art for researchers, practitioners, policymakers, fun-

This manuscript has been accepted for presentation at ***AGENTICS 2026***, the International Conference on Agentic and Generative Techniques in Intelligent Computational Systems, held in Angers, France, 28-30 October 2026, and for ***publication in Springer proceedings***. This is the authors' manuscript version. The final authenticated publication will be available via Springer.

ders, and guideline developers. That privileged role explains why reporting standards such as PRISMA 2020 [29] have become foundational across disciplines: the credibility of a review depends not only on the studies it includes, but also on whether the review process itself is transparent, reproducible, and open to critical scrutiny. PRISMA provides the modern reporting anchor through a statement paper, checklist, expanded checklist, abstract checklist, and flow diagrams, while related extensions such as PRISMA-S [34], PRISMA-ScR [40], PRISMA-P [26], and SWiM [5] support search transparency, scoping reviews, protocols, and non-meta-analytic synthesis. Together, these guidelines have substantially improved review reporting, but they were not intended as a governance standard for AI-assisted review.

The review family itself has also expanded as many contemporary studies are not conventional intervention-effect reviews. Scoping reviews are used to clarify concepts, characterize bodies of evidence, and identify gaps [4]; mapping reviews and evidence-and-gap maps emphasize landscape description, categorization, and evidence distribution [17]; and software-engineering mapping studies often rely on keywording, classification schemes, and structured maps rather than pooled causal estimation [19]. The implication is methodologically important: a guideline for AI-assisted evidence synthesis cannot assume a single review product. The acceptable role of AI depends on whether the review aims to estimate effects, map a field, structure a taxonomy, monitor a living corpus, or support broader evidence-informed reflection.

The urgent need for an AI-aware guideline arises from the collision of two developments: *i)* evidence synthesis has become slower, costlier, and more difficult to keep current; *ii)* AI systems are already entering the workflow, often before robust norms for validation, logging, and disclosure have stabilized. A recent cross-disciplinary review found that only about 5% explicitly reported machine-learning use, mostly for screening, with much rarer use in search, extraction, or synthesis [39]. A recent scoping review of LLM use in systematic reviews identified 37 relevant studies covering 10 of 13 review steps; search, screening, and extraction were the most common targets, and results were mixed rather than uniformly positive [21].

There is methodological opportunity as LLM-assisted title and abstract screening has shown promising performance in some settings. In one public-policy review of opioid-related policies, GPT-4 recommended excluding 41,742 of 43,480 articles with abstracts, leaving 1,738 for manual assessment, and the validation study reported an estimated false-exclusion rate of 0.00, with an upper confidence bound of 0.05 in the final test [36]. In a living systematic review setting, refined GPT-4o prompts achieved 100% sensitivity for studies ultimately included after full-text screening, with simulated workload reductions of 65% to 85% [13]. In a three-layer psychiatric-review study, GPT-4 reached human-comparable sensitivity with very high specificity after accounting for justified exclusions and processed roughly 110 records per minute [24]. At the same time, other evaluations found only poor-to-moderate performance for several screening

and extraction tasks outside optimized settings, which is why permissive claims about *autonomous reviewing* are not justified by the present evidence base [21].

There is also methodological risk, e.g., search-string generation with LLMs can be useful for brainstorming and initial query construction, but recent evaluations show they still lack the sensitivity and precision required for unsupervised final search design [1]. Data extraction performance varies substantially across domains and variable types, and when LLMs summarize scientific research, they can overgeneralize beyond what the original studies support: research showed LLM-generated science summaries were nearly five times more likely than human summaries to contain overly broad generalizations, with some models overgeneralizing in 26% to 73% of cases [32].

That creates the central design challenge for a modern guideline paper: *how to use AI without surrendering scientific control*. Researchers [11] emphasize that evidence synthesists remain ultimately responsible for the review, that AI must be used in ways consistent with legal and ethical standards, and that transparent reporting is mandatory. The same direction is reinforced by several publishers that require or strongly recommend disclosure of AI use and reject the idea that AI systems can bear authorship responsibility.

Existing reporting standards remain essential but incomplete for AI-assisted review. PRISMA 2020 [29] instructs authors on what to report in a systematic review; PRISMA-S [34] clarifies how to report searches; PRISMA-P [26] supports protocol registration; PRISMA-ScR [40] supports scoping-review reporting; PRESS [25] structures peer review of search strategies; and SWiM [5] helps reviewers report synthesis when meta-analysis is not appropriate. While remaining indispensable, none of them answers five questions that have become central in AI-assisted review practice: *i)* when AI use is methodologically justified; *ii)* which tasks are low-risk enough for assistive automation; *iii)* which outputs require benchmark validation before they can influence the evidence base; *iv)* which decisions require mandatory human sign-off; and *v)* what exactly readers must be told if AI has shaped the search, screening, extraction, coding, or prose. ARISMA is designed to fill that gap without displacing the classical standards on which it depends.

This gap is not only methodological but ethical: AI-assisted evidence synthesis redistributes epistemic authority across reviewers, models, vendors, platforms, and infrastructures. If ungoverned, that redistribution can affect inclusion decisions, evidence visibility, interpretive framing, privacy, accountability, and the credibility of downstream policy or clinical guidance. ARISMA therefore treats AI use in evidence synthesis as an ethics-of-methods problem: the central question is not whether AI can accelerate review work, but under what governance conditions such acceleration remains transparent, accountable, privacy-respecting, and scientifically justified.

ARISMA does not treat AI as a replacement reviewer, but as an inspected, bounded, and reversible assistant. In ARISMA, the role of AI is acceptable only when four conditions are met: *i)* the task is explicitly delimited; *ii)* the model's inputs, outputs, and constraints are documented; *iii)* performance has been val-

idated against a human reference appropriate to the task; and *iv)* humans retain both override power and accountability for all consequential decisions. This framing aligns with the current direction of AI risk-management thinking, which emphasizes trustworthiness, governance, transparency, validation, and documented human responsibility rather than unqualified automation [28].

ARISMA is designed as a governance and reporting framework rather than as a performance claim about any individual model, vendor, platform, or automation architecture. The framework therefore focuses on methodological accountability, provenance, validation, auditability, and human oversight principles that remain applicable despite rapid evolution in AI ecosystems.

This paper makes four contributions:

- It offers an end-to-end taxonomy of AI roles across systematic reviews, scoping reviews, and mapping studies.
- It provides stepwise methodological guidance for each major review activity, including searching, deduplication, screening, snowballing, extraction, categorization, synthesis, and reporting.
- It adds a governance layer through explicit validation workflows, provenance tracking, audit requirements, and human checkpoints.
- It introduces practical reporting instruments, an AI-integrated checklist, validation matrix, and AI-aware evidence-flow logic that allow editors, peer reviewers, and readers to inspect where AI entered the process and whether its role remained scientifically legitimate.

In this way, ARISMA seeks to make AI-assisted reviews more auditable, more governable, and more defensible rather than making reviews *more automatic*.

The remainder of the paper proceeds from development logic to operational use. Section 2 describes the development basis and expert-informed refinement of ARISMA. Section 3 introduces the framework and operational process. Section 4 develops the governance, validation, and provenance layer. Section 5 reviews current tool support and sustainability considerations. Section 6 presents the reporting checklist and validation matrix. Section 7 discusses threats to validity, Section 8 presents declarations, and Section 9 concludes with implications for responsible AI-assisted evidence synthesis.

2 Development Basis and Expert-Informed Refinement of ARISMA

Development rationale. ARISMA was developed through a literature-informed design process that combined established evidence-synthesis guidance, recent empirical and methodological work on AI-assisted review methods, and structured expert consultation. The authors first identified methodological commitments shared across major review standards, including transparent search reporting, explicit eligibility criteria, reproducible selection procedures, traceable extraction, and defensible synthesis [29]. They then examined where AI systems were

being introduced into these activities [21]. This process produced a draft lifecycle model, an operating framework, and initial reporting and validation instruments. The preliminary framework was intentionally conservative, with emphasis on auditability, reversibility, and human accountability.

Structured expert consultation. The draft framework was reviewed through 45- to 60-minute video consultations with 21 researchers and information specialists experienced in evidence synthesis. Participants were approached purposively through recent methodological and AI-assisted evidence-synthesis publications, with additional recommendations from initial consultees. The purposive approach was selected to obtain relevant methodological expertise rather than a statistically representative sample [35]. The semi-structured consultation format was informed by problem-centered expert interviewing [9]. Discussion focused on five areas: *i)* lifecycle coverage; *ii)* the boundaries of assistive, adjudicative, and generative AI use; *iii)* the clarity of governance checkpoints; *iv)* the usability of the checklist and validation matrix; and *v)* the accuracy of the figures.

Review and use of feedback. The authors organized the consultation feedback around these five areas and used the issues raised to guide revision. The exercise was formative: it was intended to identify omissions and improve the clarity and practical usability of ARISMA. Key issues raised concerned benchmark validation, model-version and prompt logging, model instability, stopping rules, and human oversight for high-consequence tasks.

Resulting refinements. In response to the consultation, the calibration and pilot-testing requirements were expanded; model metadata and prompt documentation were added to the reporting checklist; conditions for revising or disabling AI support were clarified; and responsibilities for human review of screening, extraction, appraisal, and synthesis outputs were strengthened. These refinements were considered together with the published methodological and empirical literature, which remains the principal evidential basis for ARISMA.

Consent and confidentiality. All experts provided informed consent for participation and for the use of anonymized methodological feedback. Identifying information was removed from the consultation records used during manuscript preparation, and no identifiable quotations are reported.

3 ARISMA Framework and Operational Process

3.1 Taxonomy and AI-use categories

ARISMA organizes AI-assisted evidence synthesis into six linked phases of ***framing, protocolization, retrieval, selection and enrichment, evidence structuring,*** and ***synthesis and reporting***. A taxonomy is necessary as different review families require different evidentiary products, and AI can only be judged appropriately in relation to those products. A scoping review may legitimately prioritize breadth, coding, and gap identification; a mapping study may prioritize category schemes and frequency distributions; a systematic review of effects may require outcome-level extraction, risk-of-bias assessment, and quantitative

or SWiM-compliant synthesis. The acceptable role of AI changes across those designs [30].

ARISMA also distinguishes assistive, adjudicative, and generative AI use. Assistive use means query suggestion, deduplication, ranking, tagging, or form prepopulation. Adjudicative use means classification, eligibility judgment, or appraisal support. Generative use means drafting text, summaries, category labels, or interpretations. The higher the epistemic consequence of the task, the stricter the validation and human oversight must be. In ARISMA, assistive uses are generally easier to justify; adjudicative uses require benchmarked performance against a human reference standard; and generative uses are never accepted as evidence in themselves, only as drafts grounded in already verified review data. That hierarchy follows directly from both the positive validation studies and the growing evidence of overgeneralization and instability in unconstrained generation [32].

Figure 1 demonstrates the governing logic of ARISMA. Its center column (white colored) represents the review process, from background framing to expert validation and release of the final audit package. The left (green) column defines the points at which human control is mandatory, and the right (gray) column defines the auditable artefacts that each step must produce. The shaded AI boxes (in light blue) represent limited intervention zones whose legitimacy depends on human oversight and the audit trail left behind.

Consultation feedback highlighted the importance of reversible AI intervention points. Accordingly, the left column was labeled "Human control," and arrows were added to show that each AI box is a bounded intervention requiring mandatory sign-off.

Below, we explain how this control logic applies across the review lifecycle, identifying where AI may assist, where validation is required, and where human accountability must remain explicit.

Background. The background step should establish why a review is needed, what decision problem it serves, and why the chosen review type is the right design. For scoping and mapping reviews, the background must justify breadth, concept clarification, and gap identification; for effect reviews, it must justify a more focused causal or evaluative question. AI can assist by surfacing terminology variants, identifying adjacent literature, clustering early search results, and drafting concept maps. It should not determine the problem statement. A strong AI-assisted background section, therefore, uses AI only for exploratory breadth, then freezes the final conceptual boundaries in human-authored form. In practice, the safest pattern is to ask an LLM to generate alternative framings and likely synonym families, then compare those suggestions against sentinel papers, prior reviews, and domain-expert feedback before the protocol is finalized.

Goal framing. The next step is to decide whether the study is a systematic review, scoping review, mapping review, evidence and gap map, or hybrid design. This decision determines not only the research question but also the acceptable role of AI. For example, AI-based cluster labeling may be very useful in a mapping review, but the same functionality is insufficient for an interven-

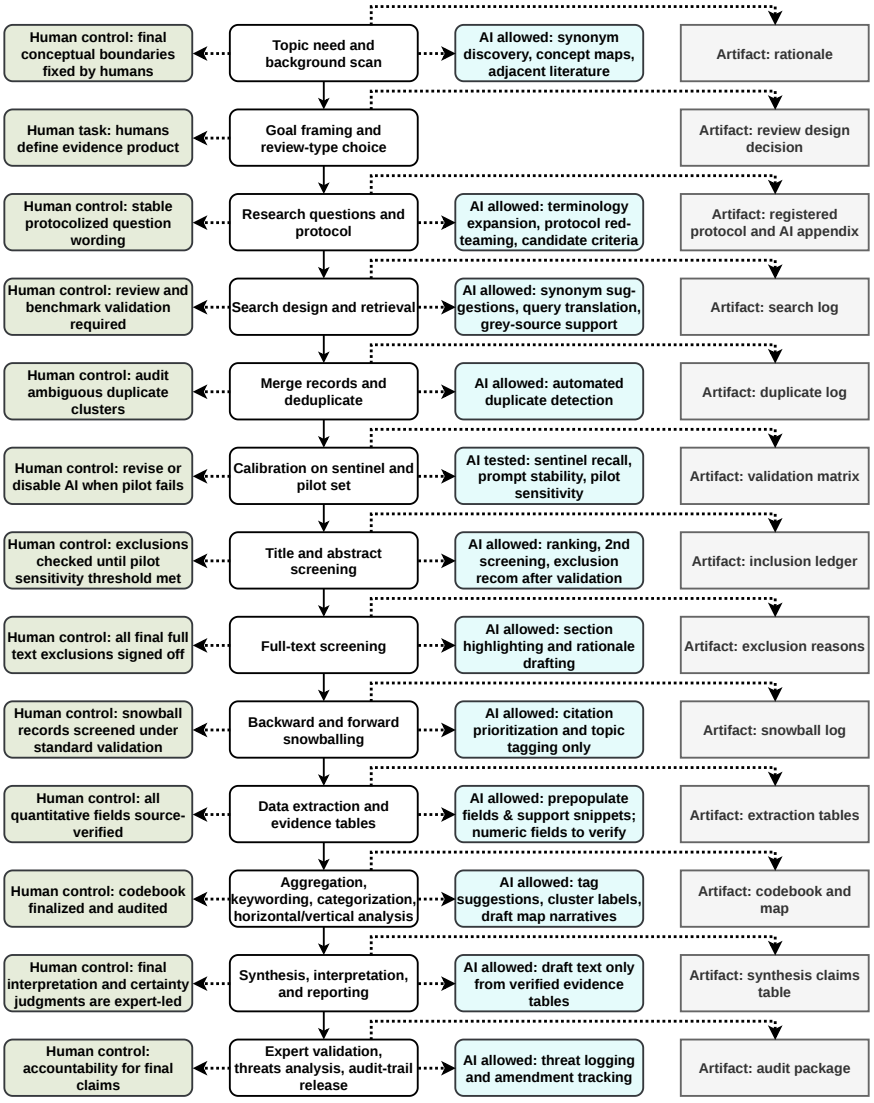

Fig. 1: General ARISMA process. AI support is attached to auditable artefacts and constrained by human validation checkpoints.

tion review that will support guideline recommendations. Goal framing should therefore specify the intended output: pooled effect estimate, narrative synthesis, evidence map, taxonomy, trend analysis, gap map, or methodological landscape. If a team cannot state the final evidence product clearly, it is too early to introduce AI into any consequential step.

Research questions. AI- and LLM-assisted reviews still begin with classical review-question logic. Population, Intervention, Comparison, Outcome (PICO) remains central for many effect reviews [10], while Population, Concept, Context (PCC) [31] and related frameworks are common in scoping reviews, and mapping studies often rely on facet-based questions such as population, method, domain, intervention family, outcome class, country, or publication venue. AI can be useful here in decomposing a broad question into searchable concept blocks, generating candidate inclusion boundaries, and stress-testing whether a question is too broad, too narrow, or internally inconsistent. However, the final question must be protocolized in stable human language because all later validation depends on it. If the review team lets question wording drift during model prompting, AI performance cannot be interpreted, and the review ceases to be auditable.

Protocolization and registration. A protocol is where ARISMA becomes conservative by design. PRISMA-P [26] exists because prospective protocols reduce selective methods drift and improve transparency, and registries such as PROSPERO exist to reduce unintended duplication and reporting bias [37]. In ARISMA, the protocol must contain not only the ordinary review methods but also an AI methods appendix stating the task, model or tool, version or build, access mode, prompt frame, input data type, validation design, human oversight rule, stopping rule, failure mode, and reporting plan. This appendix should be treated as a formal methods section, not as a software footnote. If AI use changes during the review, the amendment must be dated and justified, exactly as other protocol changes should be.

Automatic search. Search is one of the most tempting places to use LLMs because query construction is painstaking and terminology is heterogeneous. ARISMA supports AI in search only under a librarian-plus-benchmark model. The model may propose synonyms, controlled vocabulary candidates, excluded homonyms, and translations across databases. It may also help translate a validated search between platforms. But it must not be trusted as the final search architect without peer review and empirical checks. The PRESS 2015 guideline exists precisely because electronic search strategies need structured quality review [25]. Recent work on *literature search sandbox* showed that an LLM trained on 10,346 PROSPERO search queries produced searches with median sensitivity of 85% and high numbers needed to read, with librarians concluding that such queries could be useful as starting points or topic-scoping aids but not without scrutiny [1]. Other work has explicitly cautioned that ChatGPT-built review searches require stringent evaluation [2]. ARISMA search workflow, therefore, looks as follows. *First*, humans define the main concepts and a small "must-find" benchmark set. *Second*, AI expands synonyms and controlled vocabulary candi-

dates. *Third*, the search is manually curated and PRESS-reviewed. *Fourth*, the candidate strategy is tested against the benchmark set and iteratively repaired until recall is acceptable. *Fifth*, the full strategy and all AI contributions are archived. A model-generated string that misses known studies on a pilot set is not a near miss; it is a failing search strategy. That is the level of discipline required if an AI-generated search is going to influence downstream sampling.

Since retrieval is where invisible downstream bias often begins, ARISMA treats search, enrichment, deduplication, and revalidation as a single controlled subsystem rather than as disconnected technical chores. We define a retrieval and enrichment pipeline as a subsystem of automatic search presented in Figure 2.

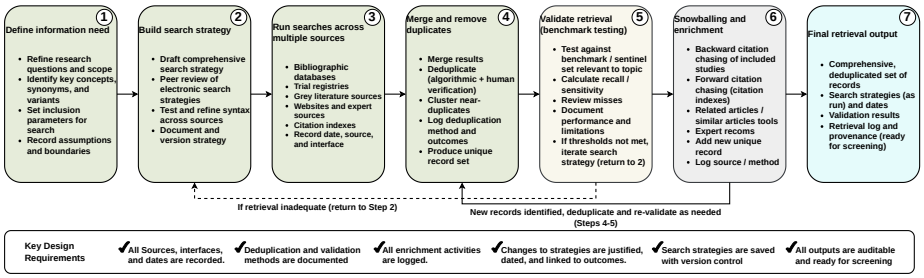

Fig. 2: ARISMA Retrieval and enrichment pipeline: search, deduplication, and snowballing.

In the defined subsystem (Figure 2), Step 1 defines the information need and the conceptual boundaries of the review. Step 2 builds and versions the search strategy. Step 3 executes the search across multiple sources and logs interfaces, dates, and source types. Step 4 merges records and removes duplicates through a combination of algorithmic matching and human verification. Step 5 validates retrieval performance against a benchmark or sentinel set. Step 6 enriches the corpus through backward and forward snowballing and related-article expansion. Step 7 produces the final retrieval output: a comprehensive, deduplicated, provenance-aware record set ready for screening. The key design requirements shown beneath the figure are the operating conditions required for auditable retrieval and adequate governance.

Consultation feedback emphasized that retrieval should be iterative rather than treated as a one-shot query event. The diagram therefore includes a loop from validation back to search, emphasized as "if thresholds are not met, iterate the search strategy," reminding teams that retrieval quality is judged by recall, transparency, and provenance.

Merging and removing duplications. Deduplication is not a mere cleaning task. Duplicate records distort workload estimates, can corrupt screening statistics, and in the worst case can lead to double counting if study reports are mishandled. ARISMA recommends that deduplication occur in two passes:

automated duplicate detection followed by human audit of ambiguous clusters and multi-report study families. Current tools discussed in Section 5 reinforce the same principle: bounded automation is more defensible than unconstrained generation. AI or ML may therefore be trusted more in duplicate detection than in free-form summary writing, but every ambiguous duplicate family still needs a human decision [12].

Inclusion and exclusion criteria framing. LLM performance in screening depends heavily on the quality and specificity of the criteria they are given. Inclusion and exclusion criteria should therefore be written as decision rules, not as aspirational prose. Criteria need explicit statements about population, exposure or intervention, outcomes, study design, language, document type, time window, publication status, and edge cases such as protocols, conference abstracts, or secondary analyses. AI is helpful here as a critic, not as the author, i.e., it can generate borderline hypotheticals, surface hidden ambiguities, and test whether different phrasings imply different judgments. Recent work in prompt development for screening and in literature prefiltering [7,6] concludes that clearer, better-structured criteria improve LLM screening behavior. ARISMA therefore treats *criteria red-teaming* as a legitimate and often useful AI contribution before actual screening begins [13].

Applying criteria during title and abstract screening. This is the most mature LLM use case, but it is not mature enough for universal autonomy. The evidence says the task can work well under the right conditions, with strong prompts, calibration, and clear asymmetry in the use case, but also that performance varies across domain, prevalence, and prompt design [7]. The public-policy GPT-4 feasibility study is persuasive for exclusion support in a setting where most records are irrelevant [36]. The living-review GPT-4o study is persuasive for prompt-refined workload reduction with continued benchmarking [13]. The three-layer GPT-4 JMIR study is persuasive for high-throughput screening under carefully engineered prompts [24]. Yet researchers show that performance remains inconsistent across tasks and languages when prompts are less optimized, or the workflow is more ambitious [18]. For that reason, ARISMA recommends a three-tier operating rule. In the *low-trust tier*, AI only ranks or labels records for human-first screening. In the *moderate-trust tier*, AI may recommend exclusions, but every exclusion is checked by a human until pilot sensitivity reaches a predeclared threshold against a consensus development set. In the *high-consequence tier*, such as reviews intended for guidelines or quantitative synthesis, AI may accelerate screening only as a second reviewer, prefilter, or prioritisation system; it must not become the sole excluding reviewer unless the team has domain-specific validation evidence and an explicit justification for doing so. This is a normative recommendation from ARISMA, grounded in the current pattern of empirical evidence rather than in any claim that a universal threshold already exists.

Full-text screening. Full-text decisions carry higher consequences than title and abstract screening because they commit the review to its final evidence base. LLMs can help here by highlighting relevant sections, extracting justifi-

cation snippets, and drafting include/exclude rationales, but the final judgment should remain a human consensus decision. The strongest current evidence suggests that full-text screening can approach human-like agreement only under constrained settings and highly reliable prompts, not as a general default [18]. In ARISMA, full-text screening is therefore a human-signoff step even if an AI assistant provides section-level reasoning or document triage. A useful practice is to require the assistant to cite page- or section-level evidence for each recommendation and to reject any full-text judgment that lacks traceable support.

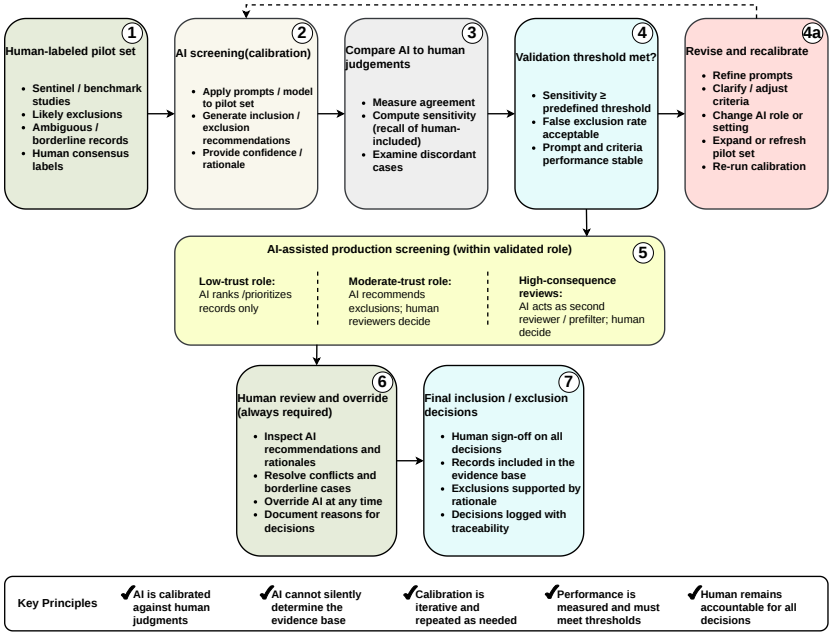

Fig. 3: Screening workflow. AI is calibrated against human judgments and cannot silently determine the evidence base.

Screening is the point at which AI can most directly and silently reshape the eventual evidence base. As shown in Figure 3, ARISMA requires task-specific calibration against a human-labeled pilot set that includes sentinel inclusions, likely exclusions, and genuinely ambiguous borderline records. Consultation feedback supported separating the workflow into a calibration phase and a production phase. The figure's top half therefore represents a calibration experiment, not a production workflow. The figure shows that AI deployment should be contingent on task-specific validation through pilot testing and that, depending on the achieved sensitivity, models can be used for ranking, recommending or second review only. Thus, AI output is first compared with human judgments, discor-

dant cases are examined, and only then is a trust tier assigned to the model's role.

The lower half of Figure 3 is a tiered-permission system. In the low-trust role, AI only ranks or prioritizes records for human-first review. In the moderate-trust role, AI may recommend exclusions, but human reviewers still decide whether those exclusions stand. In high-consequence reviews (for example, reviews intended to support formal recommendations or synthesis-critical inference) AI may at most operate as a second reviewer, prefilter, or prioritisation aid, with humans retaining the exclusion decision. The important methodological point is that these tiers are not fixed properties of any model. They are conditional permissions earned through calibration, contingent on task, corpus, and documented performance thresholds. Figure 3 is designed asymmetrically: moving from calibration to production requires passing explicit thresholds, but failing thresholds returns the process to revision and recalibration. That asymmetry encodes a precautionary principle. In ARISMA, the burden of proof lies with AI use, not with human reviewers who object to it.

Snowballing. Snowballing remains essential because database searching alone can miss semantically related but terminologically disconnected studies. Wohlin's guideline paper [41] treats backward and forward snowballing as a systematic search procedure rather than an informal afterthought and emphasizes the importance of a diverse starting set. In ARISMA, snowballing is not a merely presentational addition, as it is a controlled enrichment step that should occur after the initial included set stabilizes and again when the map or synthesis suggests missing clusters. AI can help prioritize the snowballed pool by similarity, topic tagging, or study-design detection, but it should not silently filter the snowball yield without the same validation discipline used in ordinary screening.

Calibration. Calibration is the key bridge between human review method and AI-assisted review method. It should happen before screening, before extraction, before categorization, and again when model prompts or settings change. A proper calibration exercise includes a pilot subset, blinded or semi-blinded human decisions, AI output comparison, adjudication of disagreements, and a protocolized revision of criteria or prompts. The living systematic review prompt-development study is particularly relevant [13] because it shows that strong performance did not come from a single prompt typed once into a chatbot; it came from structured development, testing, refinement, and consistency checks. ARISMA therefore treats calibration as a recurring methodological stage rather than a one-off preliminary check. Since hosted models can change without notice and corpora drift over time, ARISMA specifies *event-triggered* recalibration rather than a fixed universal interval. Recalibration on the sentinel and pilot sets is required whenever *i)* the model, version, interface, or decoding settings change; *ii)* prompts, eligibility criteria, or the codebook are amended; *iii)* corpus composition shifts materially, for example when a new database is added, a living-review update cycle begins, or retrieval volume grows beyond a predeclared proportion; or *iv)* monitoring detects sentinel misses, unstable outputs, or a rising human-override rate. In living reviews, a small sentinel benchmark

should additionally be re-run at every scheduled update cycle and logged in the validation matrix. Fixed time-based intervals may supplement, but never replace, these trigger conditions.

Language and coverage audit. AI assistance can skew evidence visibility toward English-language and terminologically mainstream literature: LLM screening performance varies across languages [18], and model-expanded queries may favor dominant vocabularies. ARISMA encourages research teams to report the language distribution of records at identification and after AI-influenced filtering steps, together with any language restrictions or translation workflows and their likely effect on coverage. Where non-English studies are eligible and resources allow, a fuller audit is a useful option: including non-English records in the sentinel and pilot sets, and reporting screening sensitivity stratified by language. If a marked asymmetry appears between the pre-AI and post-AI language distributions, it is worth investigating and documenting, and may warrant revised prompts, criteria, or human rescreening.

Data extraction. Extraction is where many teams overestimate LLM maturity. LLMs are often good at pulling salient textual descriptions, eligibility rationales, intervention names, or author-reported conclusions. They are much less reliable when the target is a numerically exact datum embedded in prose, tables, figure labels, or complex study designs. A rapid feasibility study of GPT-4 for systematic-review extraction reported around 80% overall accuracy, with better performance in some domains than others and specific difficulty with causal-inference methods and study design [38]. More recent evaluation of AI-assisted extraction from randomized clinical trials showed a sharper contrast: for binary outcomes, some models achieved high accuracy for group sizes, but event counts were only moderate, and continuous-data elements such as means and standard deviations were poor, with accuracies as low as 24% to 56% depending on the model and variable [42]. A more favorable evaluation reached a compatible conclusion: benchmarked against a reference standard set by two independent human extractors, ChatGPT-4o showed high validity and reproducibility and was judged adequate as a *second* rater alongside a human reviewer rather than as a sole extractor [27]. Taken together, these results strongly support a conservative rule: *AI may prepopulate extraction forms, but every numeric datum used in aggregation or synthesis must be checked against the source by a human reviewer.*

ARISMA extraction workflow separates fields into three classes. *Descriptive* fields such as country, study design, population label, or intervention family may be pre-extracted by AI and then quickly verified. *Interpretive* fields such as mechanism, implementation barrier, or thematic category may be AI-suggested but require human coding and reconciliation. *Critical quantitative* fields such as sample sizes, event counts, effect estimates, confidence intervals, follow-up times, and risk-of-bias support text require line-by-line or page-level human confirmation. The more the extracted datum affects a pooled estimate or a certainty judgment, the less acceptable the unverified AI extraction becomes.

Data aggregation and result gathering. ARISMA distinguishes between *result gathering* and *data aggregation* because teams often merge them prematurely. Result gathering means assembling the verified evidence objects: included studies, linked reports, extraction forms, appraisal sheets, and justifications. Data aggregation means transforming those verified objects into evidence tables, grouped datasets, map cells, or synthesis-ready frames. AI can help normalize terminology, harmonize author labels, identify likely multiple-report clusters, or draft study profile cards, but aggregation must remain tied to a provenance-aware data structure. If a synthesized statement cannot be traced back to a verified extraction field, it does not belong in the review. This is especially important in LLM-assisted workflows because generative prose can sound definitive even when it is not grounded in the extracted dataset.

Categorization and keywording. Mapping studies have long emphasized keywording and classification scheme construction as central steps; for example, keywording of abstracts is described as a way to iteratively build the categories that will structure the map [33]. LLMs can make this step much faster by proposing candidate tags, facet names, cluster labels, and latent-topic groupings from titles and abstracts. But those proposals are never the classification scheme itself. Under ARISMA, keywording is exploratory; categorization is confirmatory. The model may suggest the first, but the team must finalize the second through human consolidation, codebook definition, and back-application checks. A useful operational pattern is to let the model generate candidate keywords from a pilot set, then have two human reviewers collapse those into a controlled codebook and test inter-rater consistency before coding the full corpus.

Horizontal and vertical analysis. These steps are especially useful in scoping, mapping, and evidence-gap work. ARISMA defines *horizontal analysis* as the cross-corpus view: distributions across years, geographies, study designs, data sources, topics, interventions, outcomes, or populations. It defines *vertical analysis* as the within-cluster view: a deeper examination of what a particular category or theme actually contains, how methods differ inside it, and what findings or tensions characterize it. AI is valuable in both layers, but in different ways. For horizontal analysis, AI can generate candidate facet combinations and visual summaries from structured data. For vertical analysis, AI can assist with within-cluster thematic condensation, especially if the source material has already been coded and verified. What it must not do is invent cross-tab counts, infer absent categories, or narrate causal conclusions from map frequencies alone.

Synthesis. Quantitative synthesis and qualitative or narrative synthesis need separate AI rules. For quantitative synthesis, AI may help identify missing cells, check arithmetic consistency, suggest subgroup structures, or draft ordinary-language explanations of already computed results, but statistical model choice, heterogeneity interpretation, certainty assessment, and final effect interpretation remain expert obligations. For non-meta-analytic quantitative synthesis, SWiM is directly relevant because it exists to reduce the opacity of narrative or alternative synthesis methods [5]. For scoping and mapping reviews, synthesis often entails structured descriptions, thematic consolidation, and statements of impli-

cation rather than pooled estimates. In these settings, AI may help draft theme summaries or map narratives, but only from completed evidence tables and verified coding. Because LLMs can overgeneralize scientific findings, narrative synthesis is one of the places where unverified AI prose is most dangerous.

Methodological appraisal and quality assessment. ARISMA recommends an important distinction. Appraisal of review reports using validated checklists appears increasingly tractable for fine-tuned LLM support: a recent study found that a fine-tuned GPT-3.5 model achieved a mean accuracy of 96.5% and a mean kappa of 0.90 when supporting methodological-quality assessment of systematic reviews using a validated 27-item tool [23]. But appraisal of primary studies, especially nuanced risk-of-bias judgments, remains harder and should stay in a human-led workflow. In ARISMA, AI may therefore be used to pre-annotate appraisal forms and retrieve supporting text, but final risk-of-bias or certainty judgments must be reviewer-approved.

Expert validation of process and results. Consultation and expert review should be built into AI-assisted syntheses well before final proofreading. In the scoping review tradition, the literature argues that consultation can strengthen both process and product [20]. In ARISMA, expert validation has four legitimate targets: the framing and scope, the search coverage and missing-study risk, the categorization or codebook, and the plausibility of the synthesis claims. The process should be documented. Domain experts should not merely *react* to the final paper; they should be asked specific questions such as whether sentinel studies are missing, whether category labels distort field practice, whether the map overstates maturity, and whether AI-drafted text collapses meaningful distinctions.

Threat logging and amendment handling during conduct. AI-supported reviews need a living error ledger. Every important model failure should be recorded: missed sentinel paper, unstable classification, hallucinated extraction, over-broad summary, or unexplained output shift after a model update. Amendments to prompts, tools, or validation thresholds should be dated and justified. This is not bureaucracy for its own sake. It is the minimum scaffolding required to make the review reproducible enough for peer evaluation. A review team that cannot reconstruct how AI affected the workflow has not really conducted an auditable evidence synthesis.

4 Governance, Evidence, and Provenance

4.1 Governance and validation logic

The introduction of AI into evidence synthesis creates a methodological problem that classical review guidelines did not need to address explicitly: how to govern probabilistic systems whose outputs may vary across prompts, versions, contexts, and deployment settings. Conventional review methodology assumes that methodological steps are deterministic and reviewer-controlled. AI-assisted workflows challenge that assumption because model outputs may shift over time,

may behave differently across corpora, and may appear persuasive even when unsupported by the source evidence. ARISMA therefore treats AI use not as a software convenience, but as a governed methodological intervention requiring predefined validation rules, auditability, explicit accountability, and continuous oversight.

To operationalize this principle, ARISMA adopts a protocol-first governance logic. AI is not introduced into a review merely because a tool is available or performant in another study. Instead, the review team must first define the intended AI role, the boundaries of acceptable behavior, the validation criteria, the conditions under which the model may influence review decisions, and the situations in which AI support must be revised, restricted, or disabled entirely. Figure 4 summarizes this governance cycle and illustrates how validation, deployment, auditing, and reassessment are integrated throughout the lifecycle of an AI-assisted review.

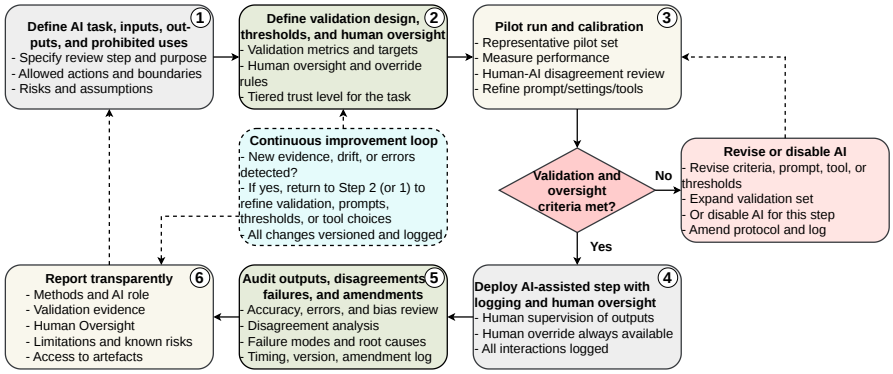

Fig. 4: AI governance and validation cycle in ARISMA: protocol first, validation deployment, and continuous oversight. Evidence synthesists remain responsible for the review. AI is an inspected, benchmarked, logged, and reversible assistant, not an autonomous reviewer.

Figure 4 operationalizes ARISMA's protocol-first governance logic. It shows that AI use is never justified by availability alone: the task, its boundaries, the validation criteria, human-override rules, and stopping conditions must be defined before deployment. The figure also makes reassessment explicit, because model suitability can change with corpus drift, prompt changes, interface updates, or newly observed failure modes.

4.2 AI-aware evidence flow and provenance

Classical evidence-flow diagrams primarily document how records move through identification, screening, eligibility assessment, and inclusion. However, once AI

systems participate in retrieval, screening, extraction, or synthesis, flow reporting alone becomes insufficient. Readers must additionally understand where AI entered the workflow, how AI outputs were validated, which decisions remained human-controlled, and how provenance was preserved from source documents to synthesis claims. In AI-assisted reviews, transparency therefore requires not only numerical accounting of records, but also methodological accounting of machine involvement.

ARISMA extends traditional PRISMA-style evidence flow by embedding provenance layers, validation checkpoints, human override rules, and audit traces directly into the evidence pipeline. The objective is to demonstrate how scientific control was maintained throughout the review rather than showing how many records were processed. Figure 5 operationalizes this AI-aware evidence-flow logic and demonstrates how auditability, traceability, and iterative enrichment are integrated into the review lifecycle.

Provenance is cumulative. Deduplication provenance records how overlapping source exports were resolved. Screening provenance records what the model recommended, what humans overrode, and how false exclusions were prevented. Extraction provenance records which fields were AI-prepopulated, which were fully human-entered, and which underwent source-level confirmation. Synthesis provenance records whether interpretive prose was derived from verified tables or generated more freely. Consultation feedback also called for a clearer chain of evidence custody showing where AI influenced records and outputs. Figure 5 therefore presents a layered provenance structure. Without that structure, readers cannot determine whether the review conclusions are anchored in reviewed evidence or have been reshaped by opaque tool behavior.

Consultation feedback supported triaging data fields according to their consequences for later synthesis. As Figure 6 shows, descriptive fields such as study setting, country, or intervention family are lower-risk and can plausibly be prepopulated by AI, subject to verification. Interpretive fields such as themes, mechanisms, barriers, or contextual labels are medium-risk because they shape later categorization and narrative synthesis; here AI suggestions may be useful, but human coders must consolidate and adjudicate. Critical quantitative fields, i.e., sample sizes, effect estimates, event counts, confidence intervals, follow-up durations, cost values, or any value that directly affects aggregation, are high-risk and should never enter synthesis without source-level human confirmation. This distinction establishes a clear boundary: AI may propose values and supporting snippets, but quantitative values and interpretive categories that feed into synthesis require human verification. This triage is also strongly supported by previously discussed literature, showing that extraction performance is uneven across field types and substantially worse for some numeric and inferential variables than for simpler descriptive content. No synthesized sentence should appear in the manuscript unless it can be traced back to a verified field, a codebook decision, or a source-confirmed analytic output.

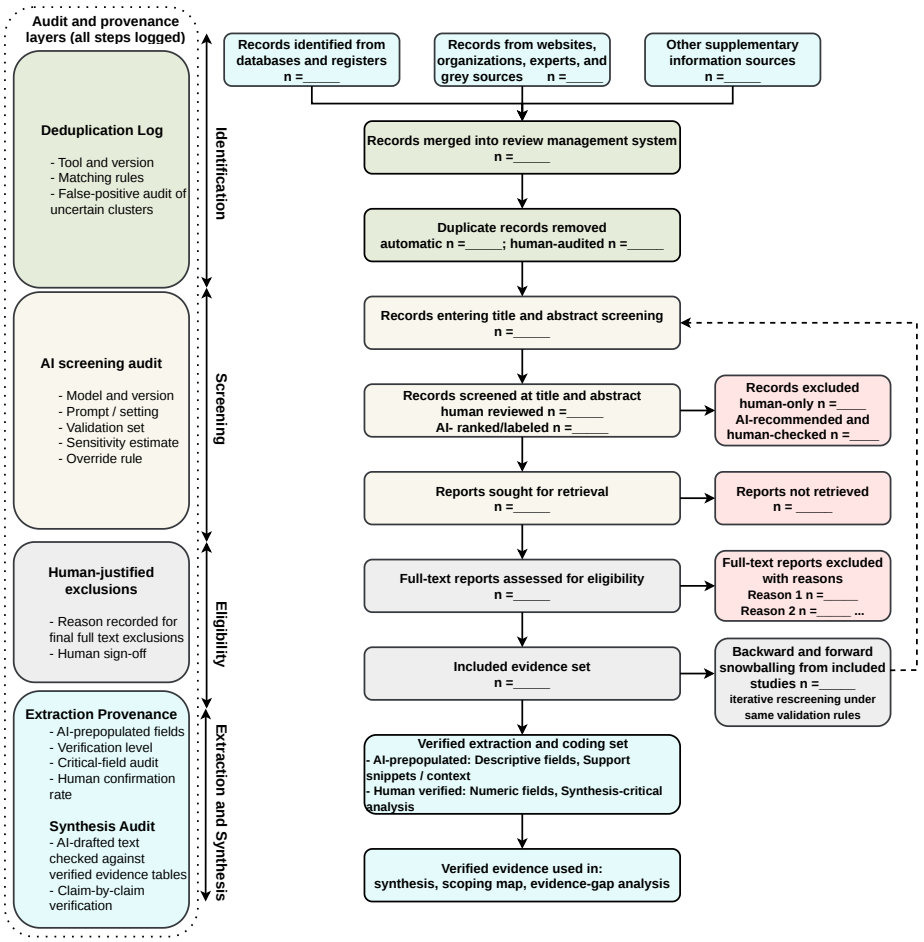

Fig. 5: ARISMA evidence flow with iterative snowballing, provenance tracking, validation, and human checkpoints. n is the number of records/reports at that step. All AI outputs are inspected, validated, and overridden by humans as needed.

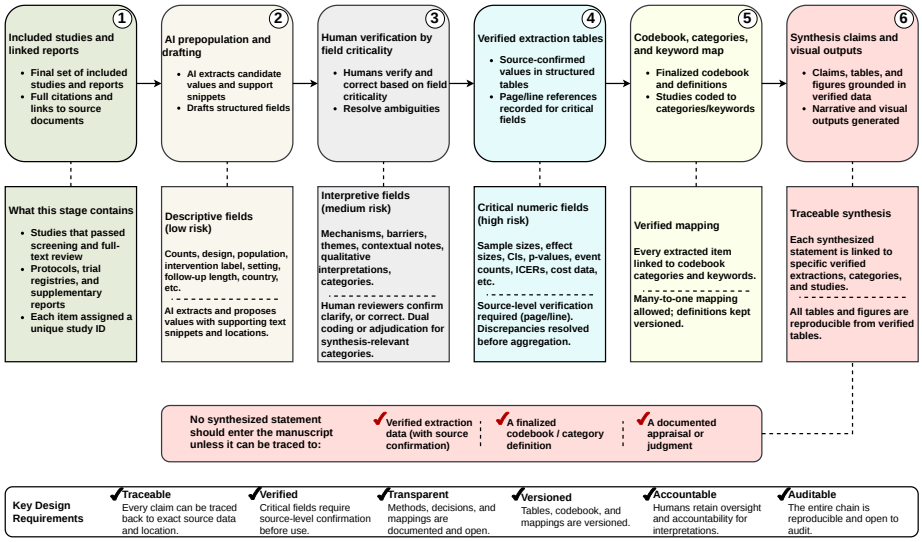

Fig. 6: Verified provenance chain from extraction to synthesis. AI can prepopulate and draft, but verified evidence tables remain the source of truth.

4.3 Legal, Privacy, and Infrastructure Governance

AI-assisted review workflows raise legal and privacy questions that classical review guidelines did not need to address explicitly. In a conventional review, bibliographic records, PDFs, extraction tables, and notes are typically processed within citation managers, spreadsheet environments, and team-shared repositories. In an LLM-assisted workflow, those same materials may be transmitted to third-party services, stored in vendor infrastructure, retained in logs, or processed across jurisdictions. The methodological consequence is that a review team can no longer treat *using a model* as a purely technical choice. It is also a data-governance choice. For ARISMA, the default rule should therefore be to classify review inputs by sensitivity before any AI use begins. Public metadata such as titles, abstracts, keywords, and database-exported bibliographic records are usually the lowest-risk inputs. Full texts, unpublished manuscripts, reviewer comments, extraction sheets containing copied passages, interview transcripts, and any documents containing personal or confidential data are higher-risk inputs and should be governed accordingly.

Three infrastructure classes should be considered:

- Public consumer interfaces, which are the least suitable route for high-stakes review work because they often provide the weakest contractual guarantees, the least predictable retention behavior from the user's perspective, and limited auditability.
- Enterprise or institutionally governed hosted interfaces, which may provide contractual, access-control, logging, or retention safeguards but still require

explicit review-team documentation of what was uploaded, for what purpose, under what access mode, and with what institutional approval.
- Local or controlled institutional deployment, including on-premise or research-cluster execution, which is often preferable for sensitive documents because it reduces exposure to external processing and gives the team greater control over versioning, logging, and retention.

Privacy assurance should be addressed around governance principles. Teams should at least articulate lawfulness, transparency, data minimization, access control, storage limitation, and integrity/confidentiality safeguards for AI-assisted processing of review materials.

Copyright and manuscript confidentiality should also be considered. Review teams often work with publisher PDFs, subscription content, or institutionally licensed materials. ARISMA recommends that authors do not upload full texts, unpublished manuscripts, or any sensitive peer-review material to external generative-AI services unless they have determined that such processing is legally and contractually permissible and consistent with institutional policy.

5 Tools and Trends

5.1 Tool Support

As detailed above, systematic reviews involve a series of labour-intensive tasks: defining scope, formulating search strategies, retrieving and deduplicating citations, screening titles and abstracts, chasing references, extracting data, assessing risk of bias, synthesizing findings and reporting. Over the last decade, dozens of tools using rule-based, machine-learning and natural language processing techniques have been developed to automate or semi-automate these tasks. ARISMA does not endorse any specific tool; rather, it classifies tools into four categories and provides governance principles for their use.

1. ***Search strategy design and translation:***
   - Visual query builders and synonym finders (e.g. 2dSearch[1]) help reviewers construct complex search strings by dragging and connecting concepts on a canvas. These interfaces improve transparency and reduce syntactic errors.
   - Search-translation utilities (e.g. Polyglot Search Translator, part of TERA [2]) convert a search string from one database syntax to another and often integrate controlled vocabularies. Search-translation utilities such as the Polyglot Search Translator can substantially reduce the manual effort required to adapt search strategies across databases while helping preserve structural consistency and recall when compared with fully manual translation workflows.

[1] `https://www.2dsearch.com/`
[2] `https://tera-tools.com/help/polyglot`

- Term-expansion tools (such as PubReMiner[3]) suggest synonyms and spelling variants; they should be used as drafting aids and not as a substitute for librarians' expertise.

2. ***Citation retrieval, deduplication and snowballing:***
   - Automated deduplicators such as ASySD[4] and the TERA[5] compare fields (title, author, DOI, year) using rule-based matching and ML. Evaluations of ASySD reported high sensitivity and specificity for duplicate detection across multiple systematic-review datasets, outperforming EndNote-based automatic deduplication approaches while substantially reducing manual review burden [12]. However, ambiguous pairs still require human checking. ARISMA therefore designates deduplication as an assistive task, with logs of removed pairs retained for audit.
   - Citation-chasing tools (e.g. Citationchaser[6], Paperfetcher[7], SpiderCite[8]) can accelerate forward and backward snowballing. Validation studies show that co-citation ranking can retrieve most eligible articles in the top ranks, saving time when more than 500 titles need to be screened [15]. Nevertheless, users must verify that the seed set is comprehensive and that relevant but infrequently cited studies are not missed. ARISMA recommends using citation-chasing tools only after the primary search has been benchmarked against sentinel studies.
3. ***Screening and prioritisation:***
   - Active-learning screeners such as Rayyan[9], Abstrackr[10], Research Screener[11], ASReview[12], SWIFT-Review [13], SWIFT-Active Screener[14], RobotAnalyst[15] and commercial platforms like DistillerSR[16] and Covidence[17] rank or classify records based on labels provided during screening. Recent evaluations of AI-assisted screening pipelines reported Work Saved over Sampling at 95% recall values ranging from approximately 49% to 87% under benchmark conditions [16]. These evaluations also emphasize the importance of rigorous calibration, stopping criteria, and continued human verification to minimize the risk of false-negative exclusions. Validation must include a pilot set, a target sensitivity (e.g. 95% recall), an

[3] https://hgserver2.amc.nl/cgi-bin/miner/miner2.cgi
[4] https://camaradesuk.github.io/ASySD/
[5] https://tera-tools.com/
[6] https://estech.shinyapps.io/citationchaser/
[7] https://paperfetcher.github.io/
[8] https://tera-tools.com/help/spidercite
[9] https://www.rayyan.ai/
[10] https://abstrackr.com/
[11] https://researchscreener.com/
[12] https://asreview.nl/
[13] https://www.sciome.com/swift-review/
[14] https://www.sciome.com/swift-activescreener/
[15] https://www.nactem.ac.uk/robotanalyst/
[16] https://www.distillersr.com/
[17] https://www.covidence.org/

explicit stopping rule (as provided by recall estimators in SWIFT-Active Screener), and documentation of all human overrides.
  - Platforms that combine screening with extraction and synthesis (e.g. Nested Knowledge[18], RevMan Web[19], JBI SUMARI[20]) enforce structured workflows and data formats but are not yet AI-driven. They can be used as organizational scaffolds around which AI-assisted tasks are inserted.
4. ***Data extraction and risk-of-bias assessment:***
  - Recent extraction and appraisal systems (e.g., AIDE[21]) can pre-populate structured fields and surface support snippets, but current studies show mixed, item-dependent performance. They are most defensible as assistive tools for descriptive fields and reviewer workflow support; synthesis-critical values and final appraisal judgments still require explicit human verification [8,3].
  - End-to-end platforms (e.g. Covidence, DistillerSR) allow customized extraction forms but do not fully automate extraction. As such, they serve as management environments rather than AI tools.

Across these tool categories, ARISMA applies the same governance rule: automation may accelerate bounded tasks, but methodological accountability remains with the review team. The required level of validation and human verification increases with the epistemic consequence of the task. Three principles apply:

- ***Benchmark and calibrate:*** Before any AI component is used in production, run a pilot on a labeled subset. Evaluate sensitivity (recall) and precision (work saved) and set thresholds. If the model fails to meet the threshold, adjust the prompt or algorithm or switch to a different tool. Calibration results should be reported in the methods section and retained for audit.
- ***Document and audit:*** All AI-driven actions must be logged, i.e., the model name and version, the date of use, key settings, the labeled pilot set, and the number of records affected. For deduplication, record pairs removed; for screening, record AI-proposed exclusions and human overrides; for extraction, record AI-filled fields and human edits. These logs enable reproducibility and support peer review.
- ***Assign human sign-off:*** AI can prioritize work but may not make final judgments on eligibility, risk of bias, or synthesis. For high-consequence tasks, at least one experienced reviewer must verify outputs. Multi-model agreement (running two independent models and comparing results) can be used to increase confidence, but disagreement must always be resolved by humans.

By following these principles, researchers can take advantage of the efficiency gains of automation while preserving scientific accountability.

[18] `https://nested-knowledge.com/`
[19] `https://revman.cochrane.org/info`
[20] `https://sumari.jbi.global/`
[21] `https://github.com/noah-schroeder/AIDE`

5.2 Sustainability and Efficiency

AI deployment depends on electricity-intensive data-center or accelerator infrastructure, and international energy analyses now treat AI as a meaningful driver of future data-center electricity demand. At the same time, inference costs are highly variable across models, hardware stacks, workloads, and serving configurations, and careful optimization can substantially reduce energy use relative to unoptimized inference pipelines. The practical implication for ARISMA is not that teams must compute exact carbon footprints for every review. Rather, teams should prefer the least resource-intensive configuration that still meets methodological requirements. In practice, that means favoring bounded task-specific tools over open-ended generation, smaller or local models for routine classification when performance is adequate, batch processing over repeated interactive prompting, and human escalation only for ambiguous or high-consequence cases.

Multi-agent debate, large reasoning models, repeated self-consistency sampling, and excessively long generative prompts can increase inference cost without proportionate methodological gain. ARISMA thus discourages *compute escalation by default*. More compute is justified only when it measurably improves validation outcomes or reduces human burden without increasing epistemic risk. So that efficiency claims can be reported rather than merely asserted, ARISMA recommends a minimal *resource-disclosure set* for every AI-assisted step: the model class, size, and version; the number of calls and records processed; total input and output tokens for API-based use; and measured energy in kWh where the team controls the hardware, for example via software-based meters. From these primitives, teams may report derived intensity ratios such as tokens per screened record or *tokens per validated record*, provided these are labeled as workload proxies rather than energy measurements. Reporting at this level is already feasible: recent screening-tool evaluations log per-record token consumption and API cost [16], and energy analyses of AI inference provide the reference points needed to interpret such disclosures [14,22].

6 Reporting and Validation Checklist

Tables 1 and 2 convert ARISMA from a general framework into a submission-ready reporting and validation instrument. Table 1 specifies what AI use must be disclosed; Table 2 specifies what must be validated before AI output can influence the review.

6.1 ARISMA Item Checklist

Table 1 is a reporting instrument that prevents *AI-assisted* from becoming an empty label. In conventional methods sections, authors can sometimes mention automation in a single sentence, for example, by saying that a model assisted with screening or drafting. Under ARISMA, that is insufficient. Readers need to know what system was used, on what input, for which task, under which

Table 1: Reporting items and additional AI/LLM reporting requirements

Item	What should be reported	Additional AI and LLM reporting required
Title	Identify the review type clearly	State if the review was AI-assisted
Abstract	Structured summary of rationale, methods, results, and conclusions	Name the AI-supported steps and the level of human oversight
Rationale	Why the review was needed	Why AI was considered necessary or useful for this workflow
Objectives	Explicit review question and scope	Which steps AI was allowed to influence and which were human-only
Review design	Systematic, scoping, mapping, EGM, rapid, or other	Why the chosen AI role is appropriate for that review family
Eligibility criteria	Full inclusion and exclusion criteria	How criteria were translated into prompts, rules, or labels
Information sources	Databases, registers, websites, grey sources, handsearching	Whether AI assisted source discovery, query iteration, or web searching
Search strategy	Reproducible search strings and filters	Prompt text or logic used to generate candidate terms; whether PRESS or equivalent review was done
Record management	Import, normalization, and deduplication methods	Tool used for deduplication; human checks for ambiguous duplicates
Selection process	Who screened what, in how many stages, with what consensus process	Validation set, sensitivity target, exclusion rule, override rule, whether AI acted as ranker, second screener, or exclusion recommender, and the language distribution of records before and after AI-influenced filtering
Supplementary searching	Snowballing, citation chasing, handsearching, expert contact	Whether AI prioritized or tagged supplementary-search yields
Data items	What variables were extracted	Which variables were AI-prepopulated and which were human-only
Extraction process	Form design, pilot testing, reviewer arrangement	Model, version, prompt, format constraints, provenance capture, and human verification rate
Appraisal or quality assessment	Tool used and how judgments were reached	Whether AI pre-annotated items or support text; who finalized judgments
Synthesis methods	Grouping, aggregation, analysis, visualization, and uncertainty handling	Whether AI drafted themes, clusters, or narrative summaries, and how those drafts were checked
Results of search and selection	Counts and flow of records	Whether AI changed the candidate pool and, if so, how validation was demonstrated
Results of extraction and appraisal	Study characteristics and quality findings	Error rates or discrepancy rates between AI outputs and human-verified values
Results of synthesis	Main findings, heterogeneity, themes, gaps, or map structures	Whether any claim in the synthesis originated from AI drafting and how it was linked to verified evidence tables
Limitations	Review limitations and evidence-base limits	Known AI limitations, model instability, domain mismatch, unresolved error modes, and any observed differences in performance or coverage across languages
Registration and protocol	Registry, protocol access, amendments	AI methods appendix location and all AI-related amendments
Funding and conflicts	Financial and non-financial disclosures	Vendor relationships, paid subscriptions, API access, model licenses, and data-governance constraints
Availability	Data, code, extraction forms, appendices	Prompts, model metadata, validation scripts, and audit logs insofar as legal and ethical constraints allow

restrictions, with what validation evidence, and with what human override rule. The checklist therefore complements PRISMA-style reporting by making AI use legible at the same level of transparency expected for search strategies, eligibility criteria, screening procedures, or synthesis methods. Consultation feedback indicated that this level of detail makes AI use more inspectable and allows the checklist to function as an operational reporting aid rather than only as a disclosure list.

6.2 Validation Matrix

The validation matrix translates ARISMA's principles into step-specific evidentiary thresholds. The consultation reinforced the principle that the burden of proof should rest on the AI-assisted step: failing a validation test returns the process to calibration. The matrix therefore distinguishes between passing a threshold (a permission to proceed) and routine operation (which remains under human supervision). Those permissions are also revocable rather than permanent, which is why the matrix closes with a recalibration row: a threshold passed under one model version, prompt, or corpus does not license continued use once any of them changes.

Table 2: Validation requirements and minimum ARISMA evidence

Step	**What must be validated**	**Minimum ARISMA evidence**
Search-term generation	Recall of known or sentinel studies	Pilot benchmark set, documented failures, revised search after peer review
Deduplication	False positives and false negatives	Tool log plus human audit of uncertain duplicate clusters
Title and abstract screening	Sensitivity for relevant records	Human-coded pilot set, error review, explicit rule for handling AI exclusions
Full-text screening	Justification quality and false exclusions	Human consensus sign-off on all final exclusions
Extraction of descriptive fields	Agreement with source papers	Random or full audit, depending on field criticality
Extraction of numeric fields	Exactness of synthesis-critical values	Full human verification of all values used in aggregation or meta-analysis
Coding, keywording, categorization	Stability of code assignment	Human-reviewed codebook plus audit of assigned clusters
Appraisal support	Agreement with validated human judgments	Item-level comparison and reviewer sign-off
Narrative drafting	Fidelity to verified evidence tables	Sentence-level verification for all claims that affect conclusions
Final synthesis	No unsupported or AI-invented claims	Senior-methodologist review and domain-expert plausibility check
Recalibration (all AI-assisted steps)	Continued validity after model, prompt, criteria, or corpus change	Re-run sentinel benchmark on each trigger event; dated amendment log linking the trigger to the recalibration result

Table 2 functions to stop teams from describing an AI-supported review as rigorous when the underlying tool was never meaningfully tested for the role it played. Each row therefore answers three questions: what can go wrong at this step, what exactly must be validated before AI output is trusted, and what minimum audit trail should remain available after the review is completed. For search-term generation, the relevant concern is recall of known studies. For deduplication, it is the balance of false positives and false negatives. For screening, it is sensitivity for relevant records and the handling of unjustified exclusions. For extraction, it is agreement with the source and source-level confirmation of synthesis-critical values. For drafting, it is fidelity to verified evidence tables and the absence of AI-invented claims. For recalibration, it is whether a previously granted permission still holds after the model, prompt, criteria, or corpus has changed, evidenced by a re-run sentinel benchmark and a dated amendment log.

The matrix also helps reviewers and editors. Editors can use it to judge whether an AI-assisted paper remains under human scientific control. Peer reviewers can use it to determine whether the authors' methodological claims are supported by validation artefacts rather than by confidence in a model brand. In other words, Table 2 turns ARISMA into usable methodology at submission and peer-review time, not only during review conduct.

7 Threats to Validity

Design threat. The first threat is choosing the wrong type of review and then trying to compensate with technology. AI cannot rescue a poorly framed evidence-synthesis design. If a team really needs a focused causal answer, a vague scoping review with impressive clustering graphics is still the wrong product. Conversely, if a field is conceptually fragmented and immature, a false impression of precision produced by premature quantitative synthesis is also a design error. AI can intensify either mistake by making output appear more polished than it is.

Retrieval threat. Search incompleteness remains one of the most serious threats in all evidence syntheses. LLM-assisted query generation is not yet reliable enough to replace expert search design, and even strong systems can miss terminologically unusual but relevant studies. That is why ARISMA insists on benchmark sets, PRESS-style review, and complementary methods such as citation chasing. Snowballing itself brings another threat: citation databases differ in coverage, and forward-chasing yields are database-dependent.

Screening threat. Screening performance is context-sensitive. The encouraging GPT-4 and GPT-4o studies were conducted under constrained conditions, with refined prompts, particular prevalence structures, and explicit benchmarking. That does not mean their performance will transfer unchanged to a different domain, language mix, or eligibility regime. Automation complacency is therefore a major risk: teams may remember the success story but forget the validation conditions that made it possible. Language coverage is part of the same threat: if models screen or rank non-English records less reliably, the corpus can drift

toward English-only evidence without the team noticing, which is why ARISMA encourages reporting the language distribution before and after AI-influenced filtering.

Extraction and appraisal threat. Extraction errors can be subtle and high impact. Numeric inaccuracies, misread tables, confusion between arms or time points, and conflation of reported and inferred values may all survive into synthesis if extraction is not fully checked. Appraisal support tools face a similar boundary problem: they may retrieve supporting text efficiently, but nuanced methodological judgments still require domain and design knowledge. The current literature supports assistance, not abdication.

Generative reasoning threat. The most distinctive LLM risk is not formatting error but epistemic drift. Models can overgeneralize, omit qualifiers, smooth over contradictions, and write a more coherent claim than the underlying evidence actually permits. Studies show polished scientific summaries can still overstate what the original papers justify. For review manuscripts, that means AI-generated discussion text is a threat surface, not a neutral convenience [32].

Reproducibility threat. Closed models change over time, prompting interfaces hide implementation details, and even low-temperature settings do not guarantee identical outputs in every environment. Some published screening studies explicitly note that model updates could alter outcomes. For that reason, reproducibility in AI-assisted reviews requires logging the model, version or date, interface type, prompt, and key settings. Even then, exact replay may not be possible for externally hosted models. ARISMA therefore prefers reproducibility by auditability and validation evidence over an unrealistic promise of bit-for-bit reproduction.

Reporting and ethics threat. Underreporting AI use is now itself a methodological threat because readers cannot assess where automation entered the workflow, whether it influenced inclusion or interpretation, or whether confidentiality, copyright, and governance issues were handled properly. This is why the joint evidence-synthesis position statement, ICMJE, COPE, and WAME all converge on disclosure and responsibility. AI cannot take authorship responsibility; the review team does.

Framework-development limitation. The expert consultation was purposive and formative. It was intended to identify omissions and improve the practical clarity of ARISMA rather than to establish formal consensus or demonstrate the framework's effectiveness. The perspectives obtained may not represent all disciplines, review traditions, or institutional settings. Broader prospective evaluation of the framework therefore remains necessary.

Open methodological questions. Some parts of the workflow remain incompletely validated. The strongest unresolved areas are fully autonomous exclusion in high-stakes reviews, extraction of complex numeric data from tables and figures, primary-study risk-of-bias automation across designs other than standard RCT formats, and the safe use of LLMs for higher-order synthesis writing. Those are research frontiers, not mature defaults. ARISMA therefore treats them as areas for controlled experimentation, not routine deployment.

8 Declarations

Expert Consultation and Consent. The expert-informed refinement of ARISMA involved voluntary methodological consultation with researchers and information specialists experienced in evidence synthesis. All consultees provided informed consent for participation and for the use of anonymized, non-identifiable feedback in the manuscript. No identifiable participant information is reported.

AI Use and Authorship Responsibility. AI systems and large language models were discussed, analyzed, and evaluated as research subjects within this study. Any AI-assisted support used during manuscript preparation remained under full human supervision and verification. All scientific judgments, methodological decisions, interpretations, validations, and final manuscript content were reviewed, verified, and approved by the authors. No AI system is listed as an author and no AI system assumes responsibility for the work.

9 Conclusion

ARISMA proposes a governance-oriented framework for AI-assisted evidence synthesis grounded in validation, provenance, transparency, and human accountability. The framework does not treat AI systems as autonomous reviewers; instead, it defines the methodological conditions under which bounded AI assistance may become scientifically defensible across systematic reviews, scoping reviews, and mapping studies. The current evidence suggests that AI can meaningfully support retrieval, prioritisation, screening assistance, extraction support, categorization, and drafting under constrained and validated conditions. At the same time, the literature consistently shows that performance remains task-dependent, unstable across contexts, and insufficient to justify unconstrained automation of consequential review decisions. ARISMA therefore reframes AI integration as a governance problem rather than a productivity problem. The central methodological question is not whether AI can accelerate review workflows, but whether its influence on evidence selection, interpretation, and synthesis remains transparent, auditable, and scientifically controllable. The framework's contribution is practical as well as conceptual. By combining lifecycle guidance, validation logic, provenance structures, reporting requirements, and audit-oriented documentation, ARISMA provides review teams, peer reviewers, editors, and policymakers with a common basis for evaluating whether AI-assisted evidence synthesis remains under credible scientific oversight. As AI systems continue to evolve, the need for transparent methodological governance will likely become more important rather than less. ARISMA is intended as a conservative foundation for that governance: one that permits responsible methodological innovation without weakening the reproducibility, interpretability, and accountability on which evidence synthesis ultimately depends. Our current work involves developing ARISMA as an open-source tool.

References

[1] Adam, G.P., DeYoung, J., Paul, A., Saldanha, I.J., Balk, E.M., Trikalinos, T.A., Wallace, B.C.: Literature search sandbox: a large language model that generates search queries for systematic reviews. JAMIA open **7**(3), ooae098 (2024)

[2] Bencze, B., Sokolowski, A., Lee, J.H., Hermann, P., Hegedüs, T., Kozuma, W., Ikumi, R., Payer, M., Salgado-Peralvo, Á.O., Végh, D.: Comparing manual and chatgpt deep research on systematic search and selection in the pubmed database on the topic of dental implantology. International Journal of Dentistry **2025**(1), 2677641 (2025)

[3] Burns, J.K., Etherington, C., Cheng-Boivin, O., Boet, S.: Using an artificial intelligence tool can be as accurate as human assessors in level one screening for a systematic review. Health Information & Libraries Journal **41**(2), 136–148 (2024)

[4] Campbell, F., Tricco, A.C., Munn, Z., Pollock, D., Saran, A., Sutton, A., White, H., Khalil, H.: Mapping reviews, scoping reviews, and evidence and gap maps (egms): the same but different—the "big picture" review family. Systematic reviews **12**(1), 45 (2023)

[5] Campbell, M., McKenzie, J.E., Sowden, A., Katikireddi, S.V., Brennan, S.E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., et al.: Synthesis without meta-analysis (swim) in systematic reviews: reporting guideline. bmj **368** (2020)

[6] 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. Annals of Internal Medicine **178**(3), 389–401 (2025)

[7] Delgado-Chaves, F.M., Jennings, M.J., Atalaia, A., Wolff, J., Horvath, R., Mamdouh, Z.M., Baumbach, J., Baumbach, L.: Transforming literature screening: The emerging role of large language models in systematic reviews. Proceedings of the National Academy of Sciences **122**(2), e2411962122 (2025)

[8] van Dijk, S.H., Brusse-Keizer, M.G., Bucsan, C.C., van der Palen, J., Doggen, C.J., Lenferink, A.: Artificial intelligence in systematic reviews: promising when appropriately used. BMJ open **13**(7), e072254 (2023)

[9] Döringer, S.: 'the problem-centred expert interview'. combining qualitative interviewing approaches for investigating implicit expert knowledge. International journal of social research methodology **24**(3), 265–278 (2021)

[10] Eriksen, M.B., Frandsen, T.F.: The impact of patient, intervention, comparison, outcome (pico) as a search strategy tool on literature search quality: a systematic review. Journal of the Medical Library Association: JMLA **106**(4), 420 (2018)

[11] Flemyng, E., Noel-Storr, A., Macura, B., Gartlehner, G., Thomas, J., Meerpohl, J.J., Jordan, Z., Minx, J., Eisele-Metzger, A., Hamel, C., et al.: Position statement on artificial intelligence (ai) use in evidence synthesis across cochrane, the campbell collaboration, jbi, and the collaboration for environmental evidence 2025. Campbell systematic reviews **21**(4), cl2–70074 (2025)

[12] Hair, K., Bahor, Z., Macleod, M., Liao, J., Sena, E.S.: The automated systematic search deduplicator (asysd): a rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. BMC biology **21**(1), 189 (2023)

[13] 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 mental health **28**(1) (2025)

[14] International Energy Agency: Energy and AI. Tech. rep., International Energy Agency, Paris (2025), https://www.iea.org/reports/energy-and-ai

[15] Janssens, A.C.J., Gwinn, M., Brockman, J.E., Powell, K., Goodman, M.: Novel citation-based search method for scientific literature: a validation study. BMC medical research methodology **20**(1), 25 (2020)

[16] Kataoka, Y., Banno, M., Kyo, M., Nakao, S., Sato, T., Taito, S., Takayama, T., Tsuge, T., Tsujimoto, Y., So, R., et al.: Tiab review plugin: A browser-based tool for ai-assisted title and abstract screening. arXiv preprint arXiv:2604.08602 (2026)

[17] Khalil, H., Welch, V., Grainger, M., Campbell, F.: Methodology for mapping reviews, evidence maps, and gap maps. Research Synthesis Methods **16**(5), 786–796 (2025)

[18] Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., Hadfield, K.: Can large language models replace humans in systematic reviews? evaluating gpt-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Research Synthesis Methods **15**(4), 616–626 (2024)

[19] Kitchenham, B., Brereton, P.: A systematic review of systematic review process research in software engineering. Information and software technology **55**(12), 2049–2075 (2013)

[20] Levac, D., Colquhoun, H., O'brien, K.K.: Scoping studies: advancing the methodology. Implementation science **5**(1), 69 (2010)

[21] Lieberum, J.L., Toews, M., Metzendorf, M.I., Heilmeyer, F., Siemens, W., Haverkamp, C., Böhringer, D., Meerpohl, J.J., Eisele-Metzger, A.: Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. Journal of Clinical Epidemiology **181**, 111746 (2025)

[22] Luccioni, S., Jernite, Y., Strubell, E.: Power hungry processing: Watts driving the cost of ai deployment? In: Proceedings of the 2024 ACM conference on fairness, accountability, and transparency. pp. 85–99 (2024)

[23] Marques-Cruz, M., Pinto, F., Vieira, R.J., Bognanni, A., Perestrelo, P., Gil-Mata, S., Duarte, V.H., Barbosa, J.P., Cardoso-Fernandes, A., Martinho-Dias, D., et al.: Use of artificial intelligence to support the assessment of the methodological quality of systematic reviews. Journal of Clinical Epidemiology p. 111944 (2025)

[24] 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. Journal of Medical Internet Research **26**, e52758 (2024)

[25] McGowan, J., Sampson, M., Salzwedel, D.M., Cogo, E., Foerster, V., Lefebvre, C.: Press peer review of electronic search strategies: 2015 guideline statement. Journal of clinical epidemiology **75**, 40–46 (2016)

[26] Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A.: Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. Systematic reviews **4**(1), 1 (2015)
[27] Motzfeldt Jensen, M., Brix Danielsen, M., Riis, J., Assifuah Kristjansen, K., Andersen, S., Okubo, Y., Jørgensen, M.G.: Chatgpt-4o can serve as the second rater for data extraction in systematic reviews. PLoS One **20**(1), e0313401 (2025)
[28] NIST: Artificial intelligence risk management framework (2023)
[29] 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. bmj **372** (2021)
[30] Peters, M.D., Godfrey, C.M., Khalil, H., McInerney, P., Parker, D., Soares, C.B.: Guidance for conducting systematic scoping reviews. JBI Evidence Implementation **13**(3), 141–146 (2015)
[31] Peters, M.D., Marnie, C., Tricco, A.C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C.M., Khalil, H.: Updated methodological guidance for the conduct of scoping reviews. JBI evidence synthesis **18**(10), 2119–2126 (2020)
[32] Peters, U., Chin-Yee, B.: Generalization bias in large language model summarization of scientific research. Royal Society Open Science **12**(4) (2025)
[33] Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M.: Systematic mapping studies in software engineering. In: 12th international conference on evaluation and assessment in software engineering (EASE). BCS Learning & Development (2008)
[34] Rethlefsen, M.L., Kirtley, S., Waffenschmidt, S., Ayala, A.P., Moher, D., Page, M.J., Koffel, J.B.: Prisma-s: an extension to the prisma statement for reporting literature searches in systematic reviews. Systematic reviews **10**(1), 39 (2021)
[35] Robson, C.: Real world research. John Wiley & Sons (2024)
[36] Rubinstein, M., Grant, S., Griffin, B.A., Pessar, S.C., Stein, B.D.: Using gpt-4 for title and abstract screening in a literature review of public policies: A feasibility study. Cochrane Evidence Synthesis and Methods **3**(3), e70031 (2025)
[37] Schiavo, J.H.: Prospero: an international register of systematic review protocols. Medical reference services quarterly **38**(2), 171–180 (2019)
[38] Schmidt, L., Hair, K., Graziosi, S., Campbell, F., Kapp, C., Khanteymoori, A., Craig, D., Engelbert, M., Thomas, J.: Exploring the use of a large language model for data extraction in systematic reviews: a rapid feasibility study. arXiv preprint arXiv:2405.14445 (2024)
[39] Scotti, K.L., Young, S., Gainey, M.A., Lan, H.: Artificial intelligence and automation in evidence synthesis: An investigation of methods employed in cochrane, campbell collaboration, and environmental evidence reviews. Cochrane Evidence Synthesis and Methods **3**(5), e70046 (2025)
[40] Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D., Horsley, T., Weeks, L., et al.: Prisma extension for scoping reviews (prisma-scr): checklist and explanation. Annals of internal medicine **169**(7), 467–473 (2018)
[41] Wohlin, C.: Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th international conference on evaluation and assessment in software engineering. pp. 1–10 (2014)
[42] Yisha, Z., Zou, P., Li, S., Zhang, L., Guo, L., Gu, A., Liu, G., Liu, T., Wang, X.: Assessing data extraction in randomized clinical trials with large language models. BMC Medical Research Methodology **26**(1), 33 (2026)