

REVIEW

Open access

Structured Assessment of AI-Assisted Systematic Reviews: Workflow and Rubric Insights

Krzysztof Wiśniewski^{1*}, Małgorzata Nowak¹, Tomasz Adamczyk²

Abstract

The application of Artificial Intelligence (AI) tools to assist with systematic literature reviews (SLRs) is expanding rapidly, yet researchers still lack structured frameworks for assessing and incorporating these tools into their research practices. This paper introduces and implements a rubric-based framework designed to evaluate AI toolchains across essential SLR stages, such as literature searching, screening, and data synthesis. A mixed-methods study was executed with 63 postgraduate student researchers, who represent the early-stage investigators that serve as primary adopters of these digital workflows. The results demonstrate that workflows structured around the rubric enhanced the participants' perceived utility of AI utilities, provided organized criteria for comparative analysis, and promoted greater transparency during tool selection. Although formal indicators of inter-rater reliability (such as Cohen's or Fleiss' Kappa) were not calculated due to the fact that participants did not evaluate identical benchmark sets concurrently, the consistency of the rubric was qualitatively validated through pilot evaluations, iterative updates, and structured focus group discussions. These efforts confirm interpretative alignment among users while recognizing that future validation via independent assessors is required. This research contributes a validated assessment rubric, an operational example of an AI-augmented SLR workflow, and a discussion regarding the consequences for research training, study reproducibility, and cross-disciplinary translation.

Keywords Artificial intelligence, Data synthesis, Systematic Literature Reviews (SLR), Literature searching

*Correspondence:

Krzysztof Wiśniewski

krzysztof.wisniewski@sggw.edu.pl

¹ Department of Systematic Review and AI Workflow, Faculty of Medicine, Warsaw University of Life Sciences, Warsaw, Poland

² Department of Rubric-Based Assessment and Research Methods, Faculty of Medicine, Jagiellonian University, Krakow, Poland

Introduction

A Systematic Literature Review (SLR) represents a structured, transparent, and reproducible methodology used to locate, assess, and interpret existing scholarship within a designated field [1]. Its primary objective is to address knowledge gaps and inform future investigations by consolidating and summarizing empirical evidence on a specific topic [2]. This approach to reviewing literature is highly valuable within academic scholarship, as it delivers a comprehensive overview of the available data [3]. The execution of an SLR encompasses multiple essential stages, specifically planning the review, carrying out the search and selection, and reporting the final outcomes [1]. To preserve methodological stringency and standard quality, investigators follow established procedural frameworks and reporting standards, such as the PRISMA statement [3].

Executing an SLR is an intricate undertaking that demands meticulous attention to detail, analytical reasoning, and an in-depth familiarity with the subject under investigation. Prior research by [4] emphasizes that establishing a precise research question, executing an exhaustive literature search, and performing a rigorous quality evaluation are vital to

successful SLR completion. Early scholarship by [5] highlights the difficulties associated with locating pertinent studies, particularly within emergent or cross-disciplinary areas, and recommends leveraging professional networks. More recently, [6] examined the susceptibility of conventional literature reviews to bias, reinforcing the necessity of employing systematic and rigorous procedures.

Despite the critical role SLRs play in delivering a comprehensive summary of existing knowledge within a domain, contemporary SLR methods frequently suffer from several drawbacks. First, the immense volume of published academic papers creates a major obstacle, as the total number of scholarly works printed annually continues to expand [7]. This exponential growth in literature complicates the identification, appraisal, and synthesis of all relevant papers. Second, conventional manual literature searches frequently depend on keyword matching, which can result in overlooked citations or skewed outcomes. Third, publication bias and selective reporting often generate an incomplete depiction of the cumulative evidence, because investigations yielding positive or statistically significant outcomes face a higher probability of publication than those with negative or inconclusive results [8]. To mitigate these constraints, systematic reviews must adopt progressive strategies, including machine learning approaches and automated screening software, to identify relevant papers rapidly and precisely [9].

Emerging Artificial Intelligence (AI) innovations offer a viable means to resolve these difficulties. By automating search protocols and extracting or summarizing core conclusions, AI-driven applications can accelerate the SLR timeline while improving both reproducibility and neutrality. Current investigations [10] demonstrate that machine learning models and natural language processing approaches can effectively aid in citation screening, data extraction, and evidence synthesis. Nonetheless, these utilities are frequently assessed in isolation, and there remains minimal guidance on how to integrate them successfully into an integrated, repeatable workflow.

A variety of AI-augmented SLR systems have been introduced lately [11]. Though beneficial, these platforms regularly emphasize individual steps of the review sequence rather than offering end-to-end operational pipelines, and their assessment is frequently performed on an ad hoc basis, centering on efficiency or usability claims without utilizing a unified comparative standard. Furthermore, the majority of validation studies depend on expert-driven case analyses or static benchmarks, which do not necessarily align with the practical requirements of student or novice researchers.

The present study resolves these limitations by developing and executing a rubric-based framework designed to assess AI toolchains across the entire duration of an SLR, from initial searching to final synthesis. This methodology introduces three primary innovations. First, it offers a structured evaluation rubric built upon foundational SLR protocols (such as PRISMA and Kitchenham) that permits a systematic comparison of disparate AI utilities. Second, it embeds this rubric within a workflow-centric evaluation that was tested empirically with postgraduate students, who constitute a critical demographic of SLR authors that frequently encounter barriers during tool adoption. Third, it examines the wider consequences for reproducibility, academic training, and cross-disciplinary adaptation, demonstrating how a rubric-driven assessment can cultivate both methodological discipline and AI literacy. By embedding AI toolchains within a transparent and reproducible evaluation framework, this research advances current methodological discussions on AI-assisted evidence synthesis and provides actionable strategies for investigators, teachers, and software engineers.

This project is distinctive because it provides an organized, rubric-based framework for evaluating and embedding AI applications within SLR operations, thereby addressing the current gap where most AI applications are tested separately instead of as components of an interconnected process [10, 12, 13]. Differing from previous scholarship, our empirically derived rubric enables an explicit, multi-criteria comparison of AI utilities across all SLR stages, and we provide open access to our research materials to encourage replication and future benchmarking. By situating AI toolchains within a reproducible assessment framework, this paper moves forward the methodological dialogue surrounding AI-assisted evidence synthesis and provides practical guidance for researchers, educators, and software creators.

The developed workflow was validated with a diverse group of academic participants, revealing that the AI-assisted methodology raised the users' perceived efficiency, enhanced the perceived relevance of selected papers, and boosted

overall user satisfaction. These outcomes correspond with recent qualitative data highlighting the efficiency improvements provided by AI in evidence synthesis [12, 13]. By offering a repeatable methodology and functional directions for choosing and blending AI utilities, our framework accommodates both research activities and educational programming, providing emerging scholars with essential AI literacy and offering academic institutions a model for embedding AI within SLR instructional courses and protocols [10].

The central objective of this research is to investigate—'How AI toolchains can be systematically evaluated and integrated into the workflow of systematic literature reviews'. To guide this investigation, we established three primary research questions:

RQ1: How can AI tools be evaluated in terms of their contribution to different stages of the SLR process?

RQ2: What dimensions or criteria emerge as meaningful for comparing and selecting AI tools for SLR workflows?

RQ3: How do postgraduate researchers, as representative early-stage users, perceive the utility and limitations of rubric-guided evaluation of AI toolchains?

These questions are formulated broadly because the specific assessment benchmarks cannot be predetermined. In fact, the lack of transparent, standardized benchmarks for checking AI-assisted SLRs serves as a primary driver for this research. The design of the rubric represents the actual process of translating these open questions into concrete, organized dimensions (such as efficiency, transparency, and usability). Consequently, rather than assuming specific evaluation criteria from the outset, these research questions served to steer the creation, adjustment, and deployment of the rubric.

We present the outcomes of a two-stage study involving 63 postgraduate students who utilized the proposed toolchain within authentic academic research environments. This study also details a reproducible assessment rubric and provides a publicly open dataset to facilitate subsequent investigations. Through these provisions, we aim to deliver both a methodological structure and a practical guide for scholars intending to implement AI-augmented SLR procedures.

Broad overview of the application of AI for SLR

Artificial Intelligence can substantially optimize the SLR framework within academic scholarship by automating diverse operations, thereby rendering the execution more efficient, precise, and exhaustive. AI delivers methodologies to model and infer knowledge, dynamically process textual data, and extract insights from extensive datasets, all of which are highly applicable to the critique of scientific literature [14]. Through the deployment of AI software engineered to automate systematic reviews, investigators can accelerate the search phase, making the undertaking more practical, sustainable, and economically viable. AI utilities can continuously monitor databases for emergent scholarship, facilitate evidence synthesis, and ensure that pertinent documents are integrated into the analysis to maintain its currency [15]. Furthermore, AI algorithms guarantee an enhanced mechanism for error-free data extraction and statistical computation relative to human operators, resulting in elevated precision [16]. Recent research by [17], corroborating an earlier study by [18], established that AI methodologies can augment both efficiency and quality during evidence synthesis, with the latter showing high fidelity in literature classification. Nevertheless, human verification remains indispensable when deploying these utilities. The investigation by [19] underscored the criticality of recall metrics in AI models utilized for literature screening, while the study by [20] emphasized the diverse array of AI and machine learning approaches being adopted within libraries, notably for collection management and user engagement. When information retrieval concepts are implemented correctly, SLRs possess a distinct superiority over conventional literature reviews, and the arrival of AI-driven utilities further intensifies this advantage.

Technological challenges in the application of AI for SLR

The integration of AI utilities for SLRs in academic scholarship introduces several technological bottlenecks. The research by [17] also underscores the necessity for comprehensive human verification during the execution of AI methods, indicating that these utilities have not yet achieved complete autonomy. The study by [21] highlights the capacity of AI to improve efficiency and quality within evidence synthesis, but concurrently emphasizes the obligation for exhaustive methodological reporting and the demand for deeper investigation into the challenges and hazards linked with AI. Additional inquiry is required to fully comprehend the pedagogical, ethical, and cultural dimensions of AI-augmented education. Cumulatively, these investigations imply that although AI utilities show great promise for refining the SLR sequence, their full capacity can only be attained through continuous research and technological refinement.

Literature Review

Information search: an overview of the protocol and strategies

The information search protocols and tactics deployed by human actors vary considerably depending on the environment, such as academic scholarship, operational decision-making, or web-based information retrieval. Within academic scholarship, an organized approach to searching is vital, necessitating the creation of librarian-mediated search strings for systematic reviews. This sequence entails formulating a core inquiry, developing exhaustive search combinations across multiple bibliographic databases, and having the search terminology verified by subject matter experts [22]. Within decision-making analysis, two distinct process-tracing methods—explicit information search and verbal protocols—are utilized to comprehend the information-processing strategies that subjects display prior to arriving at a determination [23]. Regarding web-based information seeking, individuals employ diverse tactics, including central or peripheral processing, derived from information processing theories.

The study by [24] observed that individuals frequently adopt a sequential approach when the information repository is deemed trustworthy, irrespective of their baseline expertise. This demonstrates a clear dependence on the perceived credibility of the source. The inquiry by [25] highlighted the integration of temporal phrases within search queries, yielding varied success, alongside the persistent difficulty of retrieving prospective data. A recent study [26] examined the optimization of user interaction with information retrieval architecture, emphasizing the requirement for human-centric collaborative planning models. Collectively, these investigations indicate that human operators deploy an assortment of strategies during information retrieval, including a progression from broad to granular inquiries, dependency on source reliability, and the application of temporal language.

Challenges with traditional SLR process

Conventional literature review frameworks are susceptible to several complications that can generate skewed or erroneous deductions. These difficulties incorporate a deficiency in suitable critical appraisal regarding study validity, treating all forms of evidence as possessing identical weight, and restricted stakeholder consultation, which culminates in a review of minimal practical utility [6]. Obstacles linked with information retrieval within SLRs for academic scholarship encompass the difficulty of articulating exact research inquiries, identifying appropriate databases, and formulating successful search phrases [27]. The execution frequently necessitates a pilot investigation to fine-tune search parameters, and the application of Boolean syntax to merge search concepts can be highly intricate. Furthermore, the inclusion and exclusion parameters must be explicitly operationalized to guarantee that relevant scholarship is aggregated. The screening sequence is routinely performed across two tiers: a baseline screening of titles and abstracts, followed by an exhaustive full-text evaluation. Additionally, the immense volume of records generated from search engines such as Google Scholar can render the retrieval phase inefficient. Incorporating information specialists within the systematic review pipeline can significantly influence the quality of the search execution, with involvements extending from advisory support on asset selection to configuring comprehensive search strategies [28]. The SLR configuration holds an explicit edge over conventional literature reviews provided that systematic review concepts are implemented precisely [29]. Systematic reviews depend on a collection of evidence-based strategies designed to maximize

methodological rigor and mitigate bias. However, despite the escalating interest in systematic reviews, conventional approaches to synthesizing literature remain prevalent in contemporary publications.

Limitations of PRISMA

The constraints of SLR frameworks such as PRISMA encompass publication bias, given that PRISMA's tracking mechanisms may not entirely eliminate this issue, and its foundational engineering for quantitative data, which renders it less applicable for qualitative inquiries or alternative evidence synthesis models [30]. PRISMA is likewise critiqued due to its lack of alignment with reviews incorporating narrative, qualitative, or mixed-methods paradigms rather than statistical procedures, as well as its profound concentration on meta-analysis, which discounts alternative synthesis methodologies [31]. Furthermore, PRISMA may not sufficiently accommodate innovative review products such as systematic maps, and it cannot be easily configured for methodologies that depend more heavily on the introductory phases of the review operation (searching and screening). Additionally, PRISMA's recommended criteria for review execution are minimal, which impacts the comprehensive nature of the ultimate review [31]. There is also a distinct requirement for clearer directives concerning the reporting of systematic reviews dedicated to outcome measurement instruments (OMIs), because certain dimensions of PRISMA criteria exhibit minimal relevance to reviews of that nature [32].

Research gaps

To mitigate the difficulties inherent in executing an SLR, there is a distinct need for a structured search protocol and operational strategy—specifically an ad-hoc, technology stack-based workflow—that leverages diverse AI-driven utilities. The present investigation assesses a proposed model for a search protocol and its corresponding toolchain using an evaluation rubric to identify the most effective combination of utilities for the required objectives. This study involved sixty-three student researchers who implemented the proposed model to perform their academic literature synthesis. Table 1 contrasts and compares the functional boundaries and primary focuses of selected AI-SLR utilities, emphasizing the historical lack of a standardized, rubric-driven comparative framework.

Table 1. Scope and focus of selected AI-SLR tools

Tool	Core Purpose	Applicable Stages	Assessment Features	Reported Limitations in Previous Studies
ASReview	Employs active learning techniques to support study screening	Screening	Evaluated based on processing efficiency and recall performance	Restricted to the screening phase and does not provide end-to-end workflow support
Rayyan	Facilitates collaborative review and screening activities	Screening	Assessed for usability and team collaboration capabilities	Primarily designed for title and abstract screening only
Covidence	Supports workflow coordination and data extraction processes	Screening, Extraction	Extensively utilized in Cochrane reviews	Proprietary platform with limited transparency regarding internal operations
RobotReviewer	Automates risk-of-bias evaluation using NLP techniques	Data Extraction	Focuses on natural language processing–driven automation	Specialized for a narrow task and lacks broad applicability across review stages
Elicit	Assists evidence identification and synthesis through LLM-based methods	Search, Synthesis	Provides innovative large language model functionalities	Still experimental and supported by limited validation evidence

Positioning of this study

This literature analysis reveals that although contemporary AI utilities deliver encouraging features, a unified framework to analyze them comparatively across the complete duration of the SLR process remains absent. This study fills this gap by constructing and implementing a rubric-based assessment framework, developed not merely as an instrument for tool comparison but also as a pedagogical resource for postgraduate students learning to navigate AI-augmented evidence synthesis.

Study Design

Several investigations are available regarding the application of AI utilities for SLRs within academic research. For example, one study [14] delivers a comprehensive survey of AI methodologies introduced over the past 15 years to assist investigators in executing structured analyses of scientific publications. That work examines currently available AI approaches, placing specific emphasis on their core objectives, input-output variables, and the degree of required human oversight. Another investigation [33] concentrates on the validation techniques implemented for practical AI systems documented in the literature. Furthermore, a systematic review [34] evaluates the use of AI for student learning assessment, while another systematic literature review [35] supplies an overview of automated essay scoring frameworks. Nonetheless, a dedicated evaluation rubric designed specifically to assess AI tools utilized for systematic literature reviews in academic scholarship does not exist.

The present investigation was carried out in two distinct phases:

Phase 1:

- **Number of AI tools:** Initially, 15 distinct AI utilities were analyzed for this investigation. Following the final evaluation ranking, the top 3 tools identified for each individual SLR stage are displayed in Table 2.
- **Period:** June 2023—December 2023. It is necessary to note that these utilities undergo continuous upgrades and receive new feature deployments. This ranking reflects the specific features available within this designated timeframe.
- **Participants:** 18 active investigators pursuing various disciplines of postgraduate qualifications participated in this phase as part of an official university course.

Table 2. Top-scoring AI tools selected for each SLR stage based on Phase 1 rubric evaluation

SLR Phase	Tools Identified for the Phase	Highest-Ranked Tool(s)
Preliminary Search	Elicit; Semantic Scholar; Paperdigest; Perplexity; Consensus.app	Elicit
Citation Network Analysis	Litmaps; Open Knowledge Maps; Connected Papers; Inciteful.xyz; Researchrabbit	Litmaps
Source Verification	scite.ai; sci-space	Scite.ai and Scispace
Evidence Synthesis	Paperpal; QuillBot; Jenni.ai	Jenni.ai and Consensus

Phase 2:

- **Number of AI tools:** 6 specific AI utilities configured into a cohesive workflow were recommended to the student cohort for this phase.

- **Period:** June—July 2024.
- **Participants:** 45 students interested in exploring AI applications for SLRs via an official certificate course.

Each operational phase targeted four primary SLR tasks:

1. Question Formulation & Initial Search
2. Seed Paper Identification & Citation Mapping
3. Validation & Contextual Assessment
4. Evidence Extraction & Synthesis

However, it must be emphasized that the core objective of this research is limited to identifying and isolating the correct body of information using AI-driven utilities throughout the SLR process. Manuscript composition and other supplementary tasks remain outside the boundaries of this project. The scope is strictly contained to leveraging AI-based tools to determine the most accurate information set. **Figure 1** illustrates the full diagrammatic layout of the Study design.

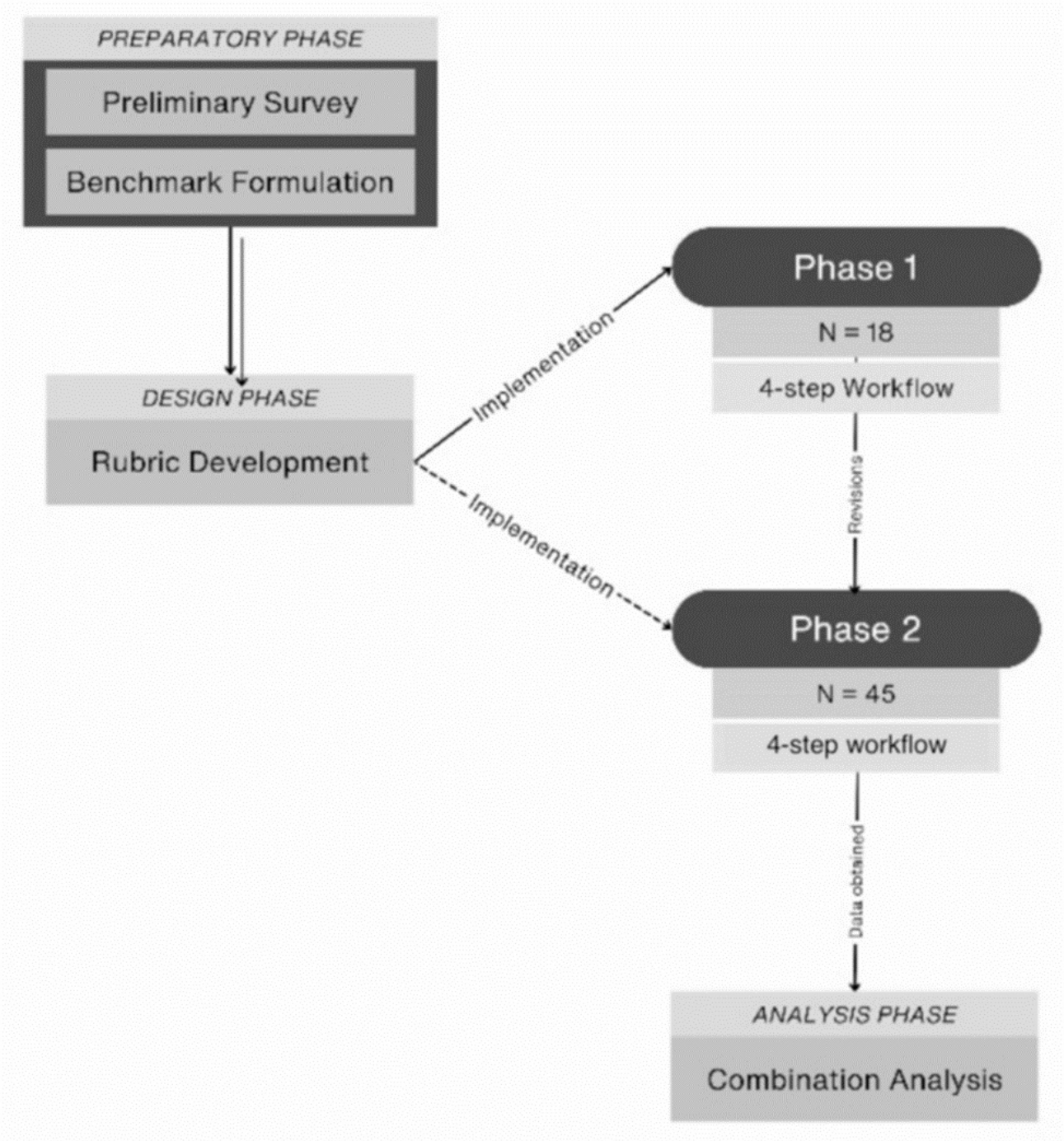


Figure 1. Study design

Developing the rubric

We constructed a multi-criteria rubric to measure the performance of each application against task-specific benchmarks (such as search precision, data visualization capabilities, and data export flexibility). Each individual benchmark was evaluated using a 3-point scale (1: Basic, 2: Moderate, 3: Advanced).

To preserve analytical objectivity:

1. Three independent evaluators graded the applications based on real-time operational testing and a thorough review of technical documentation.
2. The utilities were contrasted not merely by aggregate feature tallies, but additionally on their usability, accuracy, and educational applicability.

The evaluation rubric was updated iteratively incorporating user feedback from Phase 1, which resulted in enhanced clarity, lowered duplication, and expanded tracking of cross-tool functional capabilities.

Participant profile

The participant cohort comprised postgraduate students specializing in various academic disciplines (such as educational technology and computer science). Though they were not expert systematic reviewers, they represented the real-world academic user base most likely to adopt AI-augmented SLR practices in professional settings. All cohort members participated in a introductory 2-hour orientation session regarding AI tools for literature exploration. During Phase 2, direct workflow technical support was minimized to evaluate performance under simulated, self-directed conditions.

Workflow

Arising from the Phase 1 evaluation, the highest-rated tools isolated for each separate SLR stage were:

- **Search:** Elicit
- **Citation Mapping:** Litmaps
- **Validation:** Scite.ai and Scispace
- **Synthesis:** Consensus and Jenni.ai

This specific toolchain was implemented as the recommended operational workflow during Phase 2 and was assessed through participant scoring metrics and task completion rates.

Evaluation criteria

The performance assessment focused directly on the following dimensions:

1. **Efficiency:** Total time required to locate pertinent literature sources.
2. **Accuracy:** Relevance of the retrieved academic publications relative to predefined research topics.
3. **Usability:** User-rated simplicity and ease of interacting with each tool.
4. **Satisfaction:** The overall perception of the practical utility of the toolchain workflow.

Rubric for step 1: initial search and question formulation

This particular step appraises AI utilities based on their capacity to assist researchers in locating pertinent scholarship responding to a core research question. Four key benchmarks were established, with each evaluated via a 3-point metric.

Table 3 presents the Rubric for Step 1.

Table 3. Rubric for Step 1: Initial Search and Question Formulation

Criterion	Sub-Criterion	Score 3 (Advanced)	Score 2 (Intermediate)	Score 1 (Basic)
A. Search Efficiency	Keyword Utilization	Identifies and synthesizes relevant literature even when keyword correspondence is not exact.	Generates summaries only when exact keyword matches are present.	Unable to produce summaries based on keyword searches.
B. Tool Capabilities	Abstract Summarization	Summarizes the top N articles while providing complete metadata.	Summarizes a smaller collection of articles with only partial metadata.	Offers limited summaries or no summarization functionality.
	Filtering Flexibility	Supports multiple filtering criteria simultaneously (e.g., study type AND publication period).	Supports individual filter categories or simple conditions (e.g., OR logic).	Provides only basic filtering options (e.g., publication year).
	Accessibility	Freemium model with essential features readily available.	Free access with restricted functionality or model capabilities.	Requires payment or operates under significant access restrictions.
	Library and Export Functions	Supports categorization, storage, importing, and exporting of papers.	Allows paper storage but lacks categorization capabilities.	Does not provide saving, exporting, or sharing features.
C. Navigation	Interface Organization	Features a consistently structured interface with clear labeling and navigation.	Contains partially organized navigation with some labeling deficiencies.	Lacks navigation support for important functions.
	User-Friendliness	Provides seamless interaction with minimal learning effort.	May occasionally create confusion during use.	Presents substantial usability barriers and fragmented workflows.
D. Functional Suitability	Scholarly Literature Discovery	Delivers highly accurate retrieval of relevant academic publications.	Produces moderately relevant results that often require additional filtering.	Returns low-relevance or outdated literature.
	Concept Identification	Reveals semantic relationships and concepts across multiple studies with strong clarity.	Identifies only broad or frequently occurring themes.	Does not support concept linking or relationship discovery.
	Metadata Extraction	Accurately extracts metadata from 10–25 publications.	Metadata extraction is available but may contain limitations or occasional errors.	Metadata is frequently incomplete, inaccurate, or missing.

Every tool subjected to evaluation in Phase 1 was independently scored utilizing this rubric by three separate raters. Individual scores were subsequently averaged and normalized across all benchmarks. These values directed the final

workflow recommendations and highlighted the operational trade-offs between interface usability and functional feature density.

Rubric for step 2: seed paper identification and citation mapping

This rubric measures AI applications based on their capacity to identify foundational scholarship and visualize structural relationships connecting different literature sources. **Table 4** presents the Rubric for Step 2.

Table 4. Rubric for Step 2: Seed Paper Identification and Citation Mapping

Criterion	Sub-Criterion	Score 3 (Advanced)	Score 2 (Moderate)	Score 1 (Basic)
A. Citation Network Mapping	Graphical Representation	Provides dynamic, interactive citation-network visualizations.	Offers static visualizations or only limited interactive functionality.	Does not support citation mapping.
	Cluster Representation	Displays visual clusters derived from thematic relationships or citation density.	Supports partial clustering that often requires manual filtering.	No clustering functionality available.
	Map Search and Filtering	Enables searchable citation maps with filtering options such as author, topic, or publication year.	Includes only basic filtering capabilities.	Lacks search and filtering features within the citation map.
B. Relationship Analysis	Semantic Connectivity	Detects conceptual relationships and thematic links among publications.	Focuses primarily on citation counts or contextual citation information.	Displays only direct citation relationships without interpretation.
	Temporal Tracking	Illustrates how citation relationships evolve over time.	Provides only simple chronological sorting.	Does not include any temporal perspective.
C. System Performance	Response Time	Delivers results in < 5 s per query while maintaining stable performance.	Produces results in 5–15 s with occasional delays.	Requires > 15 s per query or experiences frequent crashes.
	Citation Reliability	Accurately reflects real-world citation networks with verifiable source tracing.	Contains some missing links or minor inconsistencies.	Includes citation links that are inaccurate or cannot be verified.
D. Search Capabilities	Multi-Modal Query Support	Accepts keyword searches, file uploads, DOI entries, and topic-based inputs.	Supports 2 out of 3 available input formats.	Requires keyword-based input only.
	Query Transparency	Clearly explains how search terms are interpreted and expanded by the system.	Provides limited insight into query-processing logic.	Offers no explanation regarding search behavior or query interpretation.

This evaluation framework places heavy emphasis on visual data processing, semantic text reasoning, and input adaptability, which are essential elements during early-stage literature discovery. The resulting scores guided the selection of Litmaps and ResearchRabbit (partially) as highly effective utilities for this operational phase.

Rubric for step 3: source validation and credibility assessment

This phase measures how effectively AI applications assist researchers in verifying the trustworthiness and situational relevance of collected academic papers. **Table 5** displays the rubric for Step 3.

Table 5. Rubric for Step 3: Source Validation and Credibility Assessment

Criterion	Sub-Criterion	Score 3 (Advanced)	Score 2 (Moderate)	Score 1 (Basic)
A. Citation Analysis Context	Citation Classification	Differentiates citations as supportive, neutral, or contradictory.	Identifies citation categories but provides little or no explanation.	Does not classify citation types.
	Contextual Citation Excerpts	Presents citation context at the sentence level.	Provides only abstract-level contextual summaries.	Does not display citation context.
B. Source Trustworthiness	Journal Standing	Displays metrics such as impact factor or Scopus ranking for the publication source.	Provides basic information about the journal.	Offers no indicators of journal credibility.
	Author Information Visibility	Links directly to author profiles and publication records.	Shows only fundamental author metadata.	Does not provide author-related information.
	Publication Recency Assessment	Identifies recent and currently influential citations.	Displays publication dates without further analysis.	Does not provide any date-based insights.
C. Annotation and Risk Assessment	User Annotation Capability	Enables users to tag, annotate, or flag potentially problematic sources.	Supports limited annotation functionality.	Does not permit annotations.
	Automated Credibility Warnings	Detects and flags potentially predatory journals or retracted articles.	Recognizes a limited number of low-trust indicators.	Provides no credibility evaluation or risk scoring.
D. Export Functionality	Citation Export and Formatting	Supports one-click export in multiple citation styles (e.g., APA, BibTeX, etc.).	Allows export in 1–2 citation formats.	Does not include citation export capabilities.

This particular rubric targets the essential question: “Can the AI tool help researchers trust the sources it retrieves?” Utilities such as Scite.ai and Scispace achieved high scores owing to their context-sensitive citation tags and integrated journal reliability indicators.

Rubric for step 4: evidence synthesis and insight extraction

This rubric evaluates tools according to their capacity to facilitate the synthesis of scholastic insights, perform perspective/stance detection, and generate high-level summaries, which are critical tasks for generating conclusions from analyzed literature. **Table 6** displays the rubric for step 4.

Table 6. Rubric for Step 4: Evidence Synthesis and Insight Extraction

Criterion	Sub-Criterion	Score 3 (Advanced)	Score 2 (Moderate)	Score 1 (Basic)
A. Summary Generation Quality	Abstractive vs. Extractive Approach	Generates coherent abstractive summaries that integrate information across multiple studies.	Extracts important phrases or passages with limited synthesis.	Relies on simple sentence aggregation or direct copy-and-paste content.
	Relevance-Based Filtering	Produces summaries tailored to user-defined topics or research questions.	Applies basic filtering using titles or keywords.	Does not perform relevance filtering during the synthesis process.
B. Knowledge Organization	Thematic Categorization	Automatically organizes findings into distinct themes, topics, or argument structures.	Groups content loosely according to keywords.	Presents findings as an unstructured list without thematic organization.
	Position and Perspective Detection	Identifies whether findings support or contradict one another.	Indicates only overall sentiment or general orientation.	Does not recognize agreement or disagreement among sources.
C. Citation Transparency	Source Attribution	Includes direct citation links within summaries to ensure traceability.	Provides only general citation references.	Does not attribute summarized content to source materials.
D. Export and Compatibility	Structured Output Export	Exports synthesized content in structured formats such as tables or outlines.	Supports export in plain-text format only.	Does not provide export functionality.
	Integration Capability	Allows seamless transfer to citation-management tools, reports, or Markdown documents.	Requires manual copying and pasting for reuse.	Generated content cannot be effectively utilized outside the platform.

Platforms like Elicit and Jenni.ai achieved the highest ratings regarding synthesis assistance due to their functionality in generating well-organized, topic-centered, and directly source-attributed summaries. Their utility in encouraging both analytical reflection and integration into academic authorship rendered them exceptionally effective during this final phase of the SLR process.

Results and Discussion

Phase 1 results: tool scoring and shortlisting

During Phase 1, 18 postgraduate investigators assessed 15 AI instruments across four fundamental steps of the SLR process utilizing an evaluation rubric. The instruments were graded on a 5-point Likert scale, and the subjects concurrently specified which system capabilities offered the highest utility. **Table 7** presents a summary of this user feedback.

Table 7. Normalized rubric scores across efficiency, transparency, and usability dimensions (n = 18, mean ± SD)

Feedback area	Percentage of participants
Rubric Score ≥ 4	77%
Perceived Search Utility	72%
Efficiency Improved	72%
Accuracy Score ≥ 4	72%
SLR process faster and with fewer errors	91%

Table 7 details the mean rubric metrics (scaled to a 5-point range) achieved by the leading tools within each sequential phase.

Note on normalization and denominators:

Every percentage displayed in **Tables 7–9** indicates the relative proportion of investigators (n = 18) who allocated a particular score within the 3-point assessment metric (1 = limited, 2 = moderate, 3 = strong). These evaluation marks were converted to a 5-point scale to allow standardized comparison across benchmarks via the linear equation:

$$N = ((S - 1) \times 2) + 1$$

where S represents the baseline evaluation score. For each designated parameter, the mean ± standard deviation (SD) across all respondents was computed. Consequently, the percentage values signify the fraction of users selecting each standardized tier; minor variances due to rounding may occur.

Table 8 displays the top-rated AI instruments selected for each individual SLR stage based on the initial Phase 1 evaluation.

Table 8. Top-scoring AI tools selected for each SLR stage based on Phase 1 rubric evaluation

SLR Phase	Highest-Rated Tool
Preliminary Literature Search	Elicit
Citation Network Analysis	Litmaps
Reference Verification	Scite.ai and Scispace
Research Synthesis	Jenni.ai and Consensus

These analytical outcomes informed the construction of a prototype AI-SLR operational pipeline for subsequent testing in Phase 2.

Phase 2 results: workflow evaluation

A total of 45 subjects evaluated the proposed AI-SLR operational workflow. Quantitative parameters (via Likert metrics) and categorical choices (concerning feature utility and educational delivery type) were compiled following the interaction. **Table 9** displays the Phase 2 outcomes of this workflow assessment.

Table 9. Weighted composite rubric scores by domain

Evaluation Dimension	Proportion of Participants (N = 45)
Rubric Rating ≥ 4	93%
Perceived Effectiveness of Search Functionality	71%
Reported Improvement in Efficiency	71%
User-Friendliness	51%
Perceived Enhancement in Accuracy	55%

Combinational analysis based on phase 2 feedback:

- **A:** Discovering and isolating relevant and related academic literature as an alternative to standard Google searches.
- **B:** Heightened efficiency while executing the SLR process.
- **C:** Simplicity and ease of operation.
- **D:** Utilizing data-driven parameters to boost result precision and lower error frequencies.

We executed a combinational analysis on the user feedback datasets to determine the specific functional preferences of the participants. **Figure 2** illustrates the outcomes of this combinational data analysis.

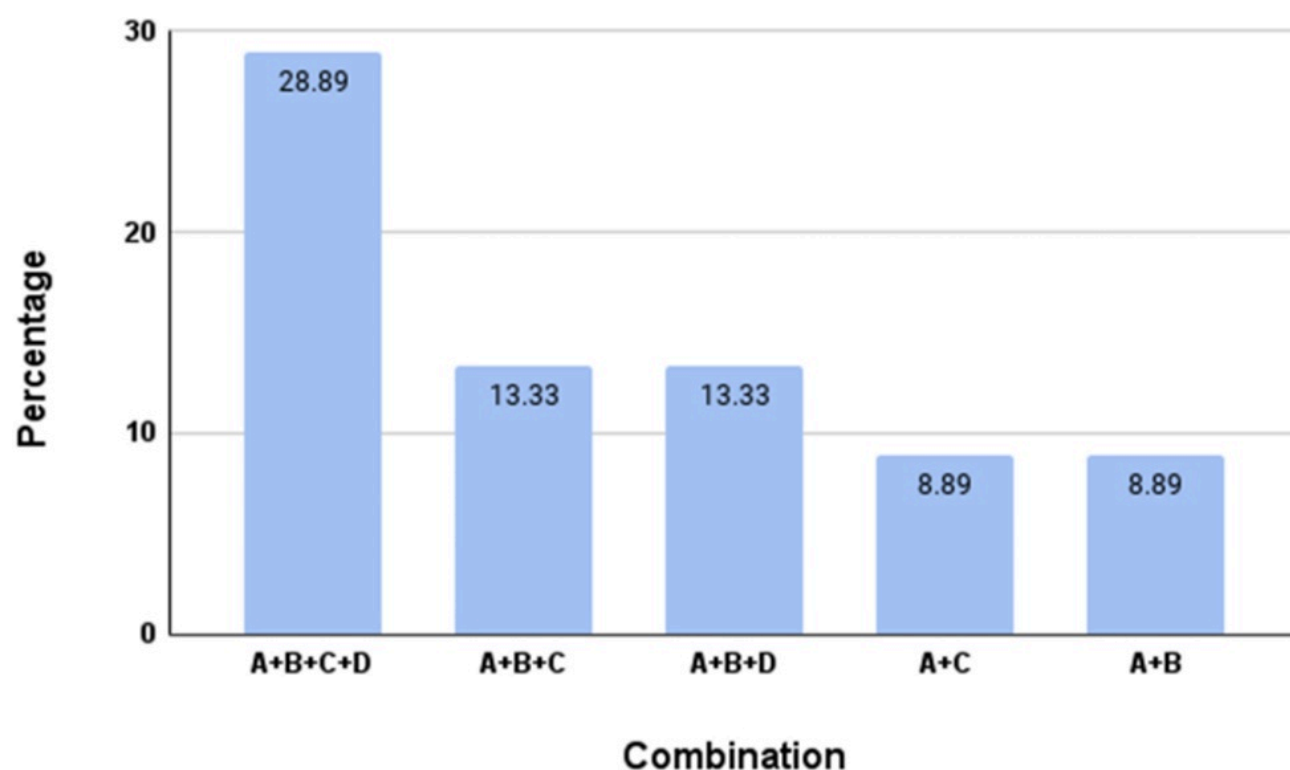


Figure 2. Combinational Analysis

Analysis based on common criteria for both phases:

The investigation gathered equivalent assessment metrics throughout both experimental steps, enabling a direct comparative analysis of user experiences with standalone instruments (Phase 1) relative to the unified operational pipeline (Phase 2). This comparison is illustrated in Table 10.

Table 10. Comparison of Phase 1 and Phase 2

Assessment Metric	Phase 2	Phase 1
Rubric Rating ≥ 4	93%	77%
Perceived Value of Search Functionality	71%	72%
Reported Efficiency Enhancement	71%	72%
Accuracy Rating ≥ 4	–	72%
Improvement in the SLR Process	–	91%
Ease of Use	51%	–
Perceived Accuracy Improvement		

Both experimental stages revealed high levels of participant satisfaction, with more than 70% of the cohort evaluating the individual tools and the unified pipeline favorably across the majority of criteria. The upward shift from 77% to 93% in cumulative rubric evaluations from Phase 1 to Phase 2 indicates that the consolidated pipeline delivered an improved user experience compared to separate utilities. User ratings for literature search utility and efficiency gains remained highly stable across both intervals (71–72%), demonstrating that the integrated pipeline sustained the advantages seen with separate utilities while expanding successfully to a larger user cohort. Conversely, the Phase 2 data indicated that only 51% of the participants classified the architecture as highly straightforward to operate, highlighting prospective targets for user interface redesign and instructional improvement.

Result analysis from participants' feedback

The qualitative and quantitative data compiled across both research stages demonstrate that the investigators found the AI-augmented SLR pipeline highly advantageous. Across both phases, roughly 71–72% of the cohort indicated that the system facilitated a more straightforward search for relevant literature and enhanced their baseline research speed. This stability suggests that the core advantages were maintained when transitioning from isolated tool assessments (Phase 1) to the execution of the comprehensive workflow (Phase 2). A substantial positive shift occurred in overall user satisfaction: during Phase 1, 77% of the cohort scored the workflow at a level of 4 or higher on a 5-point metric, whereas this metric grew to 93% during Phase 2. This advancement proves that the optimized operational pipeline performed more effectively and accommodated user requirements more successfully.

During Phase 2, we performed a closer evaluation of the functional capabilities most valued by users via a combination analysis. The subjects were permitted to choose any configuration of four primary attributes: literature searching, efficiency, simplicity, and precision. Approximately 29% of the participants selected all four attributes simultaneously, and numerous other respondents chose configurations that paired literature searching with efficiency at a minimum. This pattern indicates that while fundamental capabilities—such as targeted information retrieval and time optimization—

remain paramount, investigators also value systems that offer straightforward operation and contribute to high-fidelity outcomes. These observations provide a foundation for further refining the operational pipeline and concentrating on the system attributes that matter most to users.

Interpreting tool effectiveness

The findings validate that AI instruments can significantly optimize literature retrieval efficiency, citation networking, and data synthesis. Applications such as Elicit and Jenni.ai exhibited superior performance regarding content relevance, interface usability, and task execution speed. Visual mapping instruments like Litmaps enhanced citation path exploration, whereas Scite.ai facilitated contextual verification of source material. Participants systematically highlighted semantic functionalities (e.g., automated text summarization, query expansion) and citation tracking transparency as primary benefits. This improvement across both objective parameters (accuracy, efficiency) and subjective dimensions (user satisfaction) indicates that structured workflows constructed around curated toolchains can meaningfully alleviate the operational demands of executing SLRs.

While quantitative metrics such as precision, recall, and F_1 -score are routinely utilized within information retrieval paradigms, they cannot be directly implemented in this inquiry due to the lack of a standardized benchmark or gold-standard dataset establishing "correct" paper inclusion. Even extensively implemented frameworks such as PRISMA do not dictate an absolute ground truth against which AI application outputs can be verified. The actual relevance of academic literature within SLRs is frequently context-dependent, determined by the investigator's specific formulation of the research question, inclusion/exclusion boundaries, and disciplinary knowledge. Consequently, we prioritized participant-evaluated relevance, task speed, and rubric-driven tool performance as more alignment-appropriate metrics within this exploratory research framework.

Implications for research training

Moving beyond mere technical performance, this investigation emphasizes the pedagogical utility of structured AI workflows. Students unfamiliar with academic research practices found the evaluation rubric and selected toolchain highly beneficial for navigating intricate processes like relevance screening and text synthesis. This outcome implies that such workflows can fulfill dual functions: optimizing research outputs while simultaneously reinforcing student learning. Nonetheless, the requirement for instructional scaffolding persists. Uncritical reliance on AI-generated outputs risks diminishing deep intellectual engagement. Instructors must position AI applications as collaborative co-pilots rather than total replacements, motivating students to independently verify, cross-examine, and contextualize all generated results.

Rubric validation

Formal inter-rater reliability (IRR) coefficients, such as Cohen's or Fleiss' Kappa, were not calculated in this investigation because the evaluation framework did not involve multiple independent assessors deploying the rubric in a parallel fashion onto a unified benchmark set of tools. In point of fact, there is an explicit lack of a shared benchmark for assessing a model of this type. Each participant utilized the rubric within the unique context of their individual workflow and research objective, which fundamentally precludes the calculation of conventional IRR statistics. To establish confidence regarding the clarity and operational applicability of the rubric, we instead utilized an iterative, qualitative validation protocol: the rubric was piloted, optimized incorporating feedback from raters and participants, and distributed for structured debate among the study participants and course instructors. These procedures generated uniform interpretive themes and convergent assessments regarding tool benefits and constraints, delivering practical validation of rubric clarity while highlighting the necessity for prospective investigations to gather parallel evaluations from independent raters to facilitate formal IRR estimation.

Tool evolution and continuous sustainability

AI applications transform at a rapid pace. Functional features analyzed in this research might become obsolete or undergo corporate rebranding within a matter of months. To counter this limitation, the proposed operational pipeline prioritizes structural modularity and evaluation transparency. By ensuring that the rubric and comparative tool analyses are openly accessible, we intend to facilitate periodic re-evaluation by external researchers or academic institutions.

Threats to validity and future work

We acknowledge several structural limitations:

- **Sample limitations:** The participant pool consisted of student researchers rather than seasoned domain experts. Their assessments may not mirror the operational patterns of expert reviewers.
- **Tool selection bias:** Particular platforms (e.g., Meta, ResearchRabbit) were excluded from full analysis due to proprietary access barriers or technical instability during testing phases.
- **Short-term testing:** The evaluations were task-specific and executed over brief timeframes; long-term behavioral adoption and institutional integration were not tracked.
- **Lack of cross-domain diversity:** This study was implemented primarily within the boundaries of educational and social science literature.

Subsequent investigations should resolve these constraints by recruiting expert researchers, diversifying the academic disciplines examined, and tracking complete workflow integration within authentic research initiatives.

Conclusion

This investigation introduced and implemented a rubric-based matrix for assessing AI toolchains within the framework of systematic literature reviews (SLRs). By anchoring the rubric within validated guidelines and optimizing it through iterative participant feedback, the study demonstrated how organized benchmarks can render AI-augmented workflows more transparent, comparable, and attainable for postgraduate student researchers.

The core insights reveal that while contemporary AI utilities offer valuable assistance during specific phases—such as speed optimization during screening (ASReview), shared project management (Rayyan), or interface usability during evidence discovery (Elicit)—no single application successfully manages the complete end-to-end review lifecycle. Consequently, the rubric functions as both an analytical evaluation matrix and an educational resource, enabling researchers to balance trade-offs while advancing their overall methodological literacy.

Although formal inter-rater reliability metrics could not be computed due to the lack of a standardized benchmark set and parallel assessor scoring, qualitative validation via pilot applications and participant focus groups confirms the conceptual alignment of the rubric. Prospective research should expand this evaluation model by employing multiple independent raters and analyzing the adaptability of the rubric across alternative fields of study and varying levels of researcher expertise.

In conclusion, this study contributes a validated assessment rubric, an operational evaluation workflow model, and a discussion of the underlying educational and methodological implications. By organizing how AI toolchains are contrasted and implemented, this work charts a course toward a more transparent, reproducible, and pedagogically sound deployment of AI within evidence synthesis.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 10 Apr 2022

Revised: 12 May 2022

Accepted: 08 Jul 2022

Published online: 15 September 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

García-Peñalvo FJ. Developing robust state-of-the-art reports: systematic literature reviews. *Educ Knowl Soc.* 2022;23:e28600. <https://doi.org/10.14201/eks.28600>.

Ferreras-Fernández T, Martín-Rodero H, García-Peñalvo FJ, Merlo-Vega JA. The systematic review of literature in LIS: an approach. In: García-Peñalvo FJ, editor. *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'16)*; 2016 Nov 2-4; Salamanca, Spain. New York (NY): ACM; 2016. p. 291-8. <https://doi.org/10.1145/3012430.3012531>.

Pati D, Lorusso LN. How to write a systematic review of the literature. *HERD Health Environ Res Des J.* 2018;11:15-30. <https://doi.org/10.1177/1937586717747384>.

Anderson N, Jayaratne YSN. Methodological challenges when performing a systematic review. *Eur J Orthod.* 2015;37(3):248-50.

McManus RJ, Wilson S, Delaney BC, Fitzmaurice DA, Hyde CJ, Tobias RS, et al. Review of the usefulness of contacting other experts when conducting a literature search for systematic reviews. *BMJ.* 1998;317(7172):1562-3. <https://doi.org/10.1136/bmj.317.7172.1562>.

Haddaway NR, Bethel A, Dicks LV, Koricheva J, Macura B, Petrokofsky G, et al. Eight problems with literature reviews and how to fix them. *Nat Ecol Evol.* 2020;4(12):1582-9. <https://doi.org/10.1038/s41559-020-01295-x>.

Chu JS, Evans JA. Slowed canonical progress in large fields of science. *Proc Natl Acad Sci U S A.* 2021;118(41):e2021636118. <https://doi.org/10.1073/pnas.2021636118>.

Mlinarić A, Horvat M, Šupak Smolčić V. Dealing with the positive publication bias: why you should really publish your negative results. *Biochem Med (Zagreb)*. 2017;27(3):447-52.
<https://doi.org/10.11613/BM.2017.030201>.

van Dinter R, Tekinerdogan B, Catal C. Automation of systematic literature reviews: a systematic literature review. *Inf Softw Technol*. 2021;136:106589.
<https://doi.org/10.1016/j.infsof.2021.106589>.

Fabiano N, Gupta A, Bhambra N, Luu B, Wong S, Maaz M, et al. How to optimize the systematic review process using AI tools. *JCPP Adv*. 2024;4(2):e12234.
<https://doi.org/10.1002/jcv2.12234>.

Mikriukov A, Senokosov A, Succi G, Tormasov A, Plaksin Y, Trofimova E, et al. AI tools for automating systematic literature reviews. In: *Proceedings of the 2025 International Conference on Software Engineering and Computer Applications*; 2025. p. 25-30.
<https://doi.org/10.1145/3747912.3747962>.

Li Y, Datta S, Rastegar-Mojarad M, Lee K, Paek H, Glasgow J, et al. Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation. *J Am Med Inform Assoc*. 2025;32(4):616-25.

Mogoale PD, Pretorius AB, Mogase RC, Segooa MA. Evaluating the efficacy of AI tools in systematic literature reviews: a comprehensive analysis. *J Inf Syst Inform*. 2025;7(1):870-88.
<https://doi.org/10.51519/journalisi.v7i1.1035>.

de la Torre-López J, Ramírez A, Romero JR. Artificial intelligence to automate the systematic review of scientific literature. *Computing*. 2023;105:2171-94.
<https://doi.org/10.1007/s00607-023-01181-x>.

Atkinson CF. Cheap, quick, and rigorous: artificial intelligence and the systematic literature review. *Soc Sci Comput Rev*. 2024;42(2):376-93.
<https://doi.org/10.1177/08944393231196281>.

Collins C, Dennehy D, Conboy K, Mikalef P. Artificial intelligence in information systems research: a systematic literature review and research agenda. *Int J Inf Manage*. 2021;60:102383.
<https://doi.org/10.1016/j.ijinfomgt.2021.102383>.

Blaizot A, Veetil SK, Saidoung P, Moreno-García CF, Wiratunga N, Aceves-Martins M, et al. Using artificial intelligence methods for systematic review in health sciences: a systematic review. *Res Synth Methods*. 2022;13:353-62.
<https://doi.org/10.1002/jrsm.1553>.

Cohen AM, Ambert KH, McDonagh MS. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. *AMIA Annu Symp Proc*. 2010;2010:121-5.

Feng Y, Liang S, Zhang Y, Chen S, Wang Q, Huang T, et al. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. *J Am Med Inform Assoc*. 2022;29(8):1425-32.

Das RK, Islam MSU. Application of artificial intelligence and machine learning in libraries: a systematic review. *arXiv [Preprint]*. 2021:2112.04573.
<https://doi.org/10.48550/arXiv.2112.04573>.

Sharadgah TA, Sa'di RA. A systematic review of research on the use of artificial intelligence in English language teaching and learning (2015-2021): what are the current effects? *J Inf Technol Educ Res*. 2022;21:337-77.
<https://doi.org/10.28945/4999>.

Bramer WM, de Jonge GB, Rethlefsen ML, Mast F, Kleijnen J. A systematic approach to searching: an efficient and complete method to develop literature searches. *J Med Libr Assoc*. 2018;106(4):531-41.

<https://doi.org/10.5195/jmla.2018.283>.

Payne JW. Task complexity and contingent processing in decision making: an information search and protocol analysis. *Organ Behav Hum Perform*. 1976;16(2):366-87.

[https://doi.org/10.1016/0030-5073\(76\)90022-2](https://doi.org/10.1016/0030-5073(76)90022-2).

Sacchi S, Burigo M. Strategies in the information search process: interaction among task structure, knowledge, and source. *J Gen Psychol*. 2008;135(3):252-70.

<https://doi.org/10.3200/GENP.135.3.252-270>.

Joho H, Jatowt A, Blanco R. Temporal information searching behaviour and strategies. *Inf Process Manag*. 2015;51(6):834-50.

<https://doi.org/10.1016/j.ipm.2015.03.006>.

Schiff S, Möller R. On human-aware information seeking. In: *Humanities-Centred Artificial Intelligence, CHAI 2021: CEUR Workshop Proceedings*; 2021. p. 31-9.

Chong SW, Lin TJ, Chen Y. A methodological review of systematic literature reviews in higher education: heterogeneity and homogeneity. *Educ Res Rev*. 2022;35:100426.

<https://doi.org/10.1016/j.edurev.2021.100426>.

Eskrootchi R, Shahraki Mohammadi A, Panahi S, Zahedi R. Librarians' participation in the systematic reviews published by Iranian researchers and its impact on the quality of reporting search strategy. *Evid Based Libr Inf Pract*. 2020;15(2):69-84.

<https://doi.org/10.18438/ebliip29609>.

Mallett R, Hagen-Zanker J, Slater R, Duvendack M. The benefits and challenges of using systematic reviews in international development research. *J Dev Eff*. 2012;4(3):445-55.

<https://doi.org/10.1080/19439342.2012.711342>.

Santos MV, Morgado-Dias F, Silva TC. Oil sector and sentiment analysis—a review. *Energies*. 2023;16(12):4824.

<https://doi.org/10.3390/en16124824>.

Haddaway NR, Macura B, Whaley P, Pullin AS. ROSES reporting standards for systematic evidence syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environ Evid*. 2018;7:7.

<https://doi.org/10.1186/s13750-018-0121-7>.

Elsman EB, Butcher NJ, Mokkink LB, Terwee CB, Tricco AC, Gagnier JJ, et al. Study protocol for developing, piloting and disseminating the PRISMA-COSMIN guideline: a new reporting guideline for systematic reviews of outcome measurement instruments. *Syst Rev*. 2022;11(1):121.

Myllyaho L, Raatikainen M, Männistö T, Mikkonen T, Nurminen JK. Systematic literature review of validation methods for AI systems. *J Syst Softw*. 2021;181:111050.

<https://doi.org/10.1016/j.jss.2021.111050>.

González-Calatayud V, Prendes-Espinosa P, Roig-Vila R. Artificial intelligence for student assessment: a systematic review. *Appl Sci (Basel)*. 2021;11(12):5467.

<https://doi.org/10.3390/app11125467>.

Ramesh D, Sanampudi SK. An automated essay scoring systems: a systematic literature review. *Artif Intell Rev*. 2022;55(3):2495-527.

<https://doi.org/10.1007/s10462-021-10068-2>.