AI in education

MDPI

Article

Auditing GenAI Literature Search Workflows: A Replicable Protocol for Traceable, Accountable Retrieval in Student-Facing Inquiry

Cristo Leon [1,*] and Michelle Kudelka [2]

[1] Office of Research, Jordan Hu College of Science & Liberal Arts, New Jersey Institute of Technology, Newark, NJ 07100, USA
[2] Research, Engagement, and Access Department, Robert W. Van Houten Library, New Jersey Institute of Technology, Newark, NJ 07100, USA; michelle.kudelka@njit.edu
* Correspondence: leonc@njit.edu

Abstract

Generative AI systems increasingly mediate how students retrieve literature and generate citations, shifting methodological rigor toward the maintenance of an auditable evidence trail. This study audits the search stage of AI-assisted literature review work, focusing on retrieval performance and citation traceability rather than downstream screening or synthesis. Four widely accessible tools were compared across two retrieval postures, and Boolean queries were executed against Scopus and evaluated against a DOI-verified librarian baseline built from Scopus, Web of Science, and Google Scholar. Using a canonical prompt and a bounded top-k capture rule (k = 20), each bibliographic record was evaluated for DOI traceability, DOI resolution integrity, metadata accuracy, and run-to-run drift. Records were screened through staged title/abstract and full-text eligibility review, and the final set included 37 studies after quality appraisal was 37 studies. Across sixteen audit runs, natural-language prompting frequently produced under-target yields, recurrent integrity failures, and low overlap with the librarian benchmark. Boolean translation improved run completion and increased the proportion of auditable records, but reproducibility remained unstable across repeated runs. These findings show that correctness at the record level does not ensure stability at the evidence-set level. Limitations include the bounded tool set, the search-stage focus, and the absence of downstream screening or synthesis evaluation. Retrieval posture, therefore, emerges as a practical governance lever for AI-assisted literature review workflows and supports the use of a student-facing verification checklist anchored in DOI verification and transparent protocol capture. This research received no external funding. OSF registration: Open Science Framework, 10.17605/OSF.IO/U8NHT. The manuscript reports the final included set as n = 37, states no external funding, and lists the OSF registration DOI.

Keywords: AI tools; evidence accountability and synthesis; information retrieval; inquiry governance; large language models (LLMs); literature search; literature screening; metadata integrity; natural language processing (NLP); traceability; undergraduate and graduate research

Academic Editor: Brady D. Lund

Received: 30 January 2026
Revised: 13 March 2026
Accepted: 16 March 2026
Published: 25 March 2026

1. Introduction

Generative AI (GenAI) systems increasingly function as routine infrastructure for academic work. Undergraduate and graduate students now use conversational tools to

locate sources, draft search strategies, and generate citations. This shift moves the locus of rigor in literature search workflows from student-constructed search and screening processes to tool-mediated outputs. This study audits retrieval and citation integrity at the literature search stage; it does not evaluate downstream screening, coding, synthesis quality, or interpretive argumentation. Under these conditions, the primary risk is whether tool outputs preserve an auditable evidence trail sufficient for accountable inquiry. In evidence synthesis traditions, review quality depends on transparent search strategies, reproducible selection processes, and verifiable records (Gough et al., 2021; Grant & Booth, 2009; Snyder, 2019). In classroom settings, those same practices also function as learning outcomes. If AI outputs replace rather than support search formulation and citation checking, students can produce review-like artifacts while bypassing competencies that literature review instruction is designed to develop.

Recent scholarship frames this development as a trade-off between productivity and verifiability. Reviews of AI-assisted literature review tools report gains in speed and accessibility, alongside persistent concerns about provenance, opacity, and error propagation (Bolaños et al., 2024; Fütterer et al., 2026). Meta-level syntheses in higher education similarly call for stronger ethics, collaboration, and rigor in how AI is studied and deployed (Bond et al., 2024). Work on AI-enabled screening and workflow support demonstrates technical feasibility, but also increases the verification burden because plausible outputs can conceal bibliographic defects (Guo et al., 2024). In parallel, librarian-mediated search remains a practical benchmark for search quality and transparent reporting, with evidence that librarian involvement improves the reporting of search methods in evidence syntheses (Rethlefsen et al., 2015). Taken together, these strands motivate a governance question with direct implications for research training: how should institutions evaluate AI-assisted literature search workflows at the retrieval stage relative to an expert information-retrieval baseline, and what minimal verification practices are necessary for responsible use?

This paper addresses that question through an exploratory search-stage audit of tool-mediated retrieval and citation outputs in a student-facing context, informed by constructivist grounded theory principles for iterative protocol refinement. A librarian and a research administration practitioner observed recurring student knowledge gaps regarding the purpose and structure of literature reviews, coupled with routine deferral to AI tools for the procedural work of searching and citing. In response, the research team designed a canonical prompt to standardize tool inputs and executed paired runs across four tools representing common student access pathways: an institutionally available system: Gemini (institutional), a widely used free-tier conversational system ChatGPT (free tier; GPT-4o; hereafter, ChatGPT free tier), a paid-tier conversational system ChatGPT (paid tier; GPT-5.2 with Extended Thinking; hereafter, ChatGPT paid tier (Extended Thinking)), and an external system frequently used outside institutional environments Perplexity (Academic configuration). The audit contrasted two retrieval postures. First, natural-language prompting captured the default conversational workflow students often use. Second, each tool generated a Scopus-format Boolean query that was executed in Scopus to evaluate how database-constrained retrieval changes yield and verifiability.

The study contributes:

- a reproducible audit protocol for AI-assisted literature review workflows;
- DOI-level operational definitions for traceability and metadata integrity; and
- a student-facing governance checklist grounded in librarian practice.

The audit operationalizes traceability as DOI-based verification, evaluates metadata integrity as concordance between citation fields and authoritative source records, and estimates reproducibility as run-to-run drift under repeated execution. These outcomes translate broad concerns about AI governance into teachable and checkable indicators.

The practical endpoint is a student-facing verification checklist that supports epistemic agency: students may use AI tools, but they must demonstrate traceable, verifiable, and reproducible handling of sources rather than treating tool outputs as authoritative.

The study objectives are to: (1) quantify retrieval yield under natural-language prompting versus Boolean translation executed in Scopus; (2) operationalize and measure DOI-based traceability and major failure modes in tool-produced citations; (3) estimate run-to-run drift as an indicator of reproducibility under repeated runs; and (4) specify governance implications for classroom research workflows by anchoring search-stage tool performance to accountable inquiry expectations and an expert information-retrieval baseline. The remainder of Section 1 specifies the canonical prompt used to standardize cross-tool input, followed by a focused literature review that justifies the measures and situates the audit within evidence synthesis, information retrieval, and educational integrity debates, before detailing the reproducible protocol in Section 2.

1.1. Canonical Prompt and Standardized Cross-Tool Input

To support reproducible comparison across AI-assisted tools, the study used a single canonical prompt that fixed a topic specification derived from the study's research question, along with output constraints and reporting requirements (Box 1). The research question appears in the prompt only to standardize topical framing across tools and runs; it is not treated as a question that the tool is expected to answer authoritatively. The prompt instructed each system to: (i) return a bounded set of peer-reviewed sources relevant to the topic specification; (ii) format each citation in APA 7 and provide a DOI when available, explicitly reporting "NO DOI" when absent; (iii) provide a short relevance justification per source; (iv) generate a Scopus-format Boolean query (TITLE-ABS-KEY with parentheses and Boolean operators) for database execution; and (v) justify term inclusion and exclusion choices. This standardized input models a realistic student-facing workflow while holding constant the tasks most directly tied to traceability, metadata integrity, and reproducibility. Using a fixed prompt attributes differences in yield, integrity outcomes, and run-to-run drift to tool behavior and retrieval posture rather than to prompt variation.

Together, these objectives require a protocol that fixes tool input, constrains output, and enables repeated execution.

1.2. Literature Review

This review situates AI-assisted literature search within three linked concerns: evidentiary traceability, information-retrieval rigor, and student-facing governance. Prior work shows that AI-mediated search can produce plausible outputs while obscuring the practices needed for verification and accountability (Jesson, 2011; Kallmes et al., 2025; Post et al., 2020). Although recent syntheses describe opportunities and risks in AI-assisted review work, fewer studies operationalize verifiability as measurable integrity outcomes under realistic student-facing conditions or compare tool outputs against an information-retrieval expert baseline (Li et al., 2025; Mabirizi et al., 2025). The present audit, therefore, centers traceability, metadata integrity, and reproducibility as governance-relevant indicators aligned with evidence synthesis expectations for transparency and replicability (Eusebio et al., 2025).

The review below integrates evidence synthesis standards, information-retrieval baselines, and educational governance concerns to justify the audit constructs and verification protocol (Montoya et al., 2018).

Box 1. Canonical prompt (verbatim).

I. Prompt
You are assisting an undergraduate or graduate student conducting a literature review without librarian support.
Research question (used here as a topic anchor for retrieval standardization): How do AI-assisted tools affect the traceability, metadata integrity, and reproducibility of undergraduate and graduate literature review workflows compared to a librarian reference baseline?
Task 1: Return exactly 20 peer-reviewed sources most relevant to this question. Prefer recent and high-impact work when appropriate.
Task 2: Provide each source in APA 7 format and include a DOI for every item. If a DOI does not exist, write "NO DOI" explicitly (do not guess).
Task 3: For each source, add one short note (5–15 words) stating why it is relevant.
Task 4: Generate one Scopus-style Boolean query intended to retrieve this literature (use parentheses and field tags if appropriate).
Task 5: Provide 2–4 bullet points explaining your term inclusion/exclusion logic for the Boolean query.
Constraints: Use web browsing to verify bibliographic details when available. Do not include non-peer-reviewed sources. Do not fabricate citations or DOIs. If you are uncertain, state uncertainty and exclude the item.
II. "Allowed deviations" list (log as protocol deviations)
Tool cannot browse: record "browsing disabled/unavailable"
Tool refuses/limits citations: record "citation-limited output"
Tool cannot supply DOIs for all: mark missing DOIs explicitly and proceed
Tool cannot output Boolean: record "Boolean not supported"
III. Drift operationalization (locked to match your run plan)
Runs per tool: 2
Drift core metrics:
Jaccard similarity between Run1–Run2, Run1–Run1, Run2–Run2 (set membership of top-20 DOIs)
Percent turnover in top-20 membership across runs
Optional (if feasible): rank volatility on shared items

Note. The study research question is included only as a topical anchor to standardize retrieval behavior across tools and runs. Audit claims are derived from DOI verification and coded integrity outcomes, not from the tool's narrative response.

1.2.1. Conceptual Anchors

AI systems increasingly function as front-end interfaces for literature search and synthesis in higher education (Frey, 2018). The governance problem is straightforward: students can produce plausible bibliographic outputs without a transparent audit trail. In this study, traceability refers to whether a citation can be followed to a resolvable identifier and a matching bibliographic record. Metadata integrity refers to the correctness of the core citation fields. Reproducibility refers to whether repeated runs return a stable set of records under the same constraints (Bandara et al., 2015; Chaudhry et al., 2022; Guillen-Aguinaga et al., 2025).

1.2.2. Evidence Synthesis and Information Retrieval Baselines

Literature reviews span multiple review types with distinct rigor expectations, from narrative reviews to systematic reviews and evidence maps (Deroncele-Acosta, 2025; Gough et al., 2021). Across types, methodological guidance converges on search strategy quality, documentation, and reproducibility as core quality signals (Fink, 2019; Hernández Sampieri et al., 2014; Snyder, 2019). Work on rigor in literature reviews further emphasizes that tool support does not replace analytic judgment; instead, tools must be embedded within explicit protocols for screening, coding, and verification (Bandara et al., 2015). Empirically, librarian involvement has been associated with higher-quality reported search methods, positioning librarian practice as a pragmatic benchmark for accountable inquiry (Pawliuk et al., 2024). Recent discussions explicitly extend this baseline question to AI: whether

AI-mediated search can approximate or replace expert-mediated search, and under what transparency conditions (Park, 2025; Park et al., 2024).

1.2.3. AI-Assisted Retrieval and Writing Risks in Review Workflows

Recent syntheses and practical guides converge on a consistent message: AI can accelerate parts of review workflows, but it introduces governance risks through opacity, provenance loss, and error propagation (Bolaños et al., 2024; Bond et al., 2024; Bukar et al., 2023; Fütterer et al., 2026). Automation studies in evidence synthesis show efficiency gains while also documenting adoption barriers driven by trust, training, and workflow fit (Clark et al., 2021; Scott et al., 2021). In parallel, LLM-focused studies demonstrate feasibility for screening and other steps, while implicitly elevating traceability requirements because errors become harder to detect when outputs appear coherent (Guo et al., 2024). Across these lines of work, several failure risks are especially salient for student-facing practice: citation hallucination or fabrication, DOI mismatch (a resolving DOI attached to the wrong record), non-resolving identifiers, and ranking opacity that impedes reproducible retrieval. The emergence of reporting frameworks for AI-mediated review processes signals field recognition that documentation itself must become a primary research artifact, not an afterthought. Taken together, this work establishes verifiability and provenance loss as the dominant integrity risks in AI-assisted review workflows (Chaudhry et al., 2022; Holst et al., 2025; Park, 2025).

1.2.4. Educational Implications

In higher education, AI use raises questions of epistemic agency, assessment validity, and information literacy, particularly when students outsource procedural components of inquiry that are themselves intended learning outcomes (Basol et al., 2021; Buyya et al., 2016; Fink, 2019; Haukås et al., 2018; Lochmiller & Lester, 2016). Because AI-mediated searching and writing are already embedded in classroom practice (Baker & Caton, 2025; S. Wang et al., 2024), the institutional issue is no longer whether students use these systems, but how that use is governed. Quality control, therefore, shifts from discouraging AI use to specifying auditable requirements for source retrieval, verification, and citation handling (Hardie et al., 2025; Ndéla Marone & Mbengue, 2025).

Recent studies document classroom integration of AI tools in writing instruction, STEM learning, and project-based pedagogy, alongside variation in student perceptions, receptiveness, and ethical concerns (Höpner & Uckelmann, 2024; Kajan et al., 2025; Omeh & Ayanwale, 2025). This matters for the present audit because it indicates that AI-mediated searching and writing are already part of student workflows, often before formal information literacy instruction occurs (Pai H. et al., 2025). Under these conditions, helpfulness is not an adequate evaluation criterion. Instructional settings require governance mechanisms that protect provenance, reduce citation and identifier errors, and preserve assessment validity when AI systems mediate retrieval and bibliographic formatting (Rizky Noviandy et al., 2024; Temitope et al., 2025).

Taken together, this evidence supports treating AI-assisted literature review work as a workflow-governance problem. Yet most implementation-oriented studies still stop short of operationalizing verifiability and reproducibility at the citation and identifier level or benchmarking tool behavior against librarian-grade information retrieval practice.

1.2.5. Summary and Gap: Why "Helpfulness" Is Insufficient

Across the reviewed literature, two gaps remain consequential for student-facing practice. First, most discussions of AI in reviews focus on efficiency, usability, or broad ethical considerations, while fewer studies operationalize traceability and metadata integrity as auditable outcomes that can be verified record-by-record under a reproducible

protocol (ISO 15836, 2017; Leipzig et al., 2021). Second, direct evaluation against an expert information-retrieval baseline remains limited in educationally realistic settings, particularly where repeated runs can expose drift and instability. These gaps motivate a reproducible audit that treats tool output as data, evaluates DOI-level traceability and metadata integrity as primary outcomes, and frames governance requirements for students and institutions. Section 2 translates this gap into an operational protocol.

2. Materials and Methods

Search topic. The audit examines literature on generative-AI-assisted tools used in the literature-search stage of broader evidence-synthesis and review workflows, specifically query formulation, retrieval, and citation generation. The focal phenomenon is not synthesis quality, but search-stage auditability: whether tool outputs preserve verifiable citations and reproducible evidence sets when students use conversational systems to retrieve literature. Accordingly, the study evaluates retrieval outputs using three criteria: DOI-based traceability, metadata integrity, and run-to-run drift. These assessments are anchored to a DOI-verified, librarian-mediated database baseline consisting of Scopus, Web of Science, and Google Scholar.

Study Overview. The study implements a comparative audit of AI-assisted literature-search workflows under two retrieval postures. ChatGPT was included as an audited retrieval system in two configurations, free tier and paid tier with Extended Thinking, and was therefore analyzed as a study condition rather than used only as manuscript-support software. Four retrieval systems/configurations were evaluated: Gemini (institutional), ChatGPT free tier, Perplexity (academic configuration), and ChatGPT paid tier (Extended Thinking).

For each system, we ran two conditions: (i) a natural-language retrieval condition in which the system returned citations directly, and (ii) a Boolean-translation condition in which the system generated a Scopus-format Boolean query that was then executed in Scopus to retrieve records. To standardize inputs across systems and conditions, all runs used the same canonical prompt and a bounded capture rule of top-k, where $k = 20$. The unit of analysis was the system-produced bibliographic record.

Outcomes. The outcomes were assessed across three audit dimensions. Traceability measured whether each citation could be verified through a resolvable DOI. Metadata integrity measured whether the DOI resolved to the cited work, that is, whether the DOI and article matched correctly, and whether key bibliographic fields such as author, title, venue, and year were consistent with the DOI landing page. Reproducibility was assessed through run-to-run drift, quantified using Jaccard overlap across paired runs and turnover within the captured top-k sets. The librarian-mediated database baseline of Scopus, Web of Science, and Google Scholar provided the DOI-verified reference set used to operationalize topical scope and contextualize overlap outcomes.

Organization of Section 2. Section 2 presents the audit design, eligibility criteria, information sources, operational definitions, verification procedures, and reporting disclosures (Wätzold et al., 2024). The Disclosure of Support Statement (DSS) is provided in the Supplementary Materials. Sections 2.4–2.11 define the operational variables used in Section 3.

2.1. Study Design and Protocol

This paper employed an exploratory, autoethnographic audit design analyzed through Constructivist Grounded Theory (CGT) to examine how widely accessible AI tools support—or undermine—traceable and reproducible literature search practices. A librarian and a research administrator observed recurring student knowledge gaps regarding the purposes

and conceptual structure of literature reviews namely the *why* and the *what*, alongside increasing reliance on AI tools for procedural execution, that is the *how*, without comparable engagement with information literacy support. To address this gap, the research team designed a cross-tool prompt and executed a structured set of repeated runs to generate comparable retrieval outputs. The protocol operationalized the unit of analysis as a tool-produced bibliographic record (citation string plus identifiers), then audited each record for DOI traceability and metadata integrity under manual verification. Consistent with CGT, the team iteratively compared tool outputs across runs and tools (constant comparison), refined analytic categories through memo-writing, and used targeted additions to the corpus (LIS expert records) as theoretical sampling to strengthen coverage of foundational information retrieval and search-stage governance literature that students commonly miss.

2.2. Eligibility Criteria

The retrieval stage intentionally avoided restrictive inclusion criteria in order to characterize each tool's default behavior under realistic student-like conditions, that is, default tool settings and minimally constrained prompts. Consistent with this audit objective, we retained both peer-reviewed sources and grey literature at ingestion to document source substitution and provenance drift as observable workflow risks. Grey literature was operationally defined as web pages, video recordings, reports, preprints, and blog posts. This classification allows the audit to observe provenance and substitution risks at capture while preserving methodological rigor in the final synthesis.

At the study level, records were eligible if they examined AI-assisted tools used in the literature search stage of a review workflow, rather than only for writing, screening, data extraction, synthesis, or other processes. As a protocol rule, grey literature was ingested for provenance testing but was not eligible for inclusion in the final evidence synthesis, except for one curated grey-literature item retained for its contextual value (Weaver, 2025). This exception was documented in the screening log and traceability codebook (Supplementary Material: Document S4) and is counted in the final included set.

Eligibility decisions were applied only after consolidation, using staged screening (title and abstract screening followed by full-text assessment) and subsequent quality appraisal (Sundelson et al., 2023).[1] Exclusions and reason codes are reported in the Results. The final included set after quality appraisal was n = 37, as reflected in the PRISMA flow diagram (Figure 1).

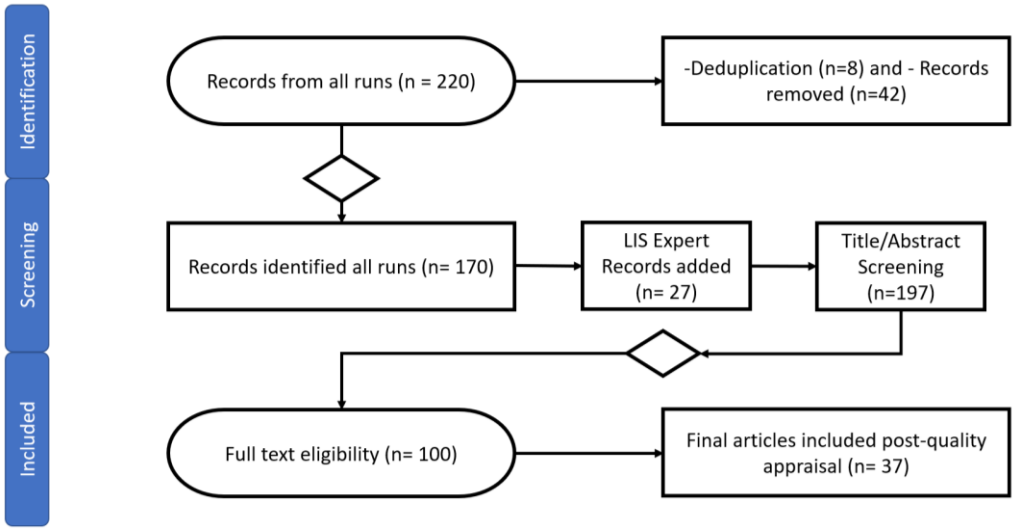

Figure 1. PRISMA flowchart: Corpus cleaning and consolidation workflow.

2.3. Information Sources

The study evaluated four AI tools reflecting tiers commonly available to students through free access, personal subscription, and institutional licensing: ChatGPT free tier, ChatGPT paid tier (Extended Thinking), Gemini (institutional) (provided via Google Workspace), and Perplexity (Academic configuration). Boolean queries generated by the tools were executed in Scopus using a consistent ranking rule ("Cited by (highest)") to standardize result selection across tools (see Appendices A.1–A.5). Web browsing was enabled during tool interactions to reduce retrieval opacity attributable to offline constraints. Records were managed and deduplicated in Zotero, then re-deduplicated using a CSV export to address residual duplicates driven by bibliographic field variation. VOSviewer, version 1.8.0_202, was used to profile the consolidated corpus via co-authorship and keyword co-occurrence overlay maps (VOSviewer, 2022).

2.4. Search Strategy

This subsection documents the retrieval protocol in two postures. First, the canonical prompt was executed as natural-language runs to observe default tool behavior under student-like prompting. Second, each tool generated a Scopus-format Boolean query that was executed in Scopus to evaluate how structured retrieval affects yield, traceability, and reproducibility. A *run* denotes one prompt execution in a tool under the specified constraints, and an *item/record* denotes one bibliographic citation captured for the audit.

2.4.1. Canonical Prompt and Natural-Language Runs

The research team executed eight natural-language runs (two per tool) using an identical canonical prompt and constraints. Each run asked the tool to return an evidence set relevant to the topic specification, and the audit's claims were evaluated via DOI verification and coded integrity outcomes rather than the tool's narrative explanation. For cross-tool comparability, extraction was limited to the first $k = 20$ citations when a run returned more than 20 items. Across the natural-language runs, the tools produced 101 outputs, including one non-result placeholder (0 positive; non-result). After removing this non-result, the natural-language posture yielded $n = 100$ items for the traceability audit.

2.4.2. Boolean Translation and Scopus Execution

For each tool, the generated Boolean query was executed in Scopus using TITLE-ABS-KEY field tags. Results were sorted by "Cited by (highest)" and truncated to the top $k = 20$ for comparability. Across the Boolean executions, the tools produced 129 outputs, including one non-result placeholder (0 positive; non-result). After removing this non-result, Scopus execution yielded $n = 128$ retrievable items. Zotero then removed 8 duplicate items during library consolidation, yielding $n = 120$ unique Boolean-stage items for the traceability audit. When a Boolean string returned zero results, the query was shortened using a pre-specified rule: the final clause segment (that is, the most specific term family appended at the end of the string) was reduced first while preserving the core construct families. In the Supplementary Materials, all query revisions are color-coded (red) to support reproducibility and auditability. Across both postures, the combined pre-consolidation corpus comprised 220 items (92 natural-language + 128 Boolean).

2.5. Selection Process and PRISMA Workflow

To transform tool-level outputs into an auditable corpus for cross-tool comparison, this subsection details a staged consolidation workflow. This approach separates tool-level yield from record-level uniqueness while preserving a reproducible audit trail. Across all tool runs, the tools produced 220 bibliographic records after removing non-result placeholders

(N/A) from the run outputs. We imported all records into Zotero and performed automated and manual deduplication, removing 8 Zotero duplicates and retaining 212 records.[2] We then exported the Zotero library to CSV and conducted a second deduplication pass to remove residual duplicates attributable to bibliographic field variation (for example, inconsistent author strings, venue name variants, and missing identifiers), yielding 170 unique records. An LIS expert subsequently contributed 27 additional records to strengthen coverage of information retrieval and evidence synthesis baselines. Following a final duplicate check, the consolidated corpus contained 197 records. Titles and abstracts were screened against the inclusion and exclusion criteria, resulting in 97 exclusions and 100 records advanced to full-text assessment. Following full-text assessment and quality appraisal, 37 studies were included in the final synthesis (full-text and/or quality appraisal exclusions, $n = 63$). All screening and adjudication decisions, including deduplication rationales and exclusion reasons, were logged in the screening log and traceability codebook.

2.6. Data Collection Process

The study generated primary data by executing and recording tool outputs under a standardized prompt protocol and then auditing the resulting bibliographic records. The team captured tool outputs, extracted citations, and entered each record into Zotero to standardize bibliographic field structure. For each record, we recorded the tool, run ID, retrieval posture (natural-language versus Boolean), citation string, DOI string (when present), and source type classification (peer-reviewed versus grey literature). The team then verified traceability using a fixed DOI audit protocol and coded each record using an operational traceability codebook.

To improve metadata integrity prior to descriptive profiling, we conducted a Zotero completion pass to harmonize fields (for example, author strings, venue titles, dates, and identifiers) and performed a secondary confirmation of keywords. Keyword integrity actions were explicitly coded in the traceability codebook using four tags: _KW-MISSING-PDF (keywords absent in PDF), _KW-VERIFIED-PUBLISHER (keywords retrieved from the journal or publisher website), _KW-SUBSTITUTED-CATEGORIES (platform categories used as a proxy for keywords in the single curated grey-literature item), and _KW-CCS-CONCEPTS (publisher-assigned ACM CCS concept labels recorded). Provenance for all keyword interventions was documented in Zotero and cross-referenced in the screening log.

In parallel, the team maintained CGT memos to document emerging failure modes (for example, DOI wrong-match and source-type substitution), support constant comparison across tools, and guide iterative refinements to the audit codebook using first-cycle coding techniques consistent with Charmaz (Charmaz, 2017) and Saldaña (Saldaña, 2021).

Student-facing verification checklist. To increase practical uptake in instructional settings, Table 1 presents a compact verification checklist derived from the fuller per-run and per-citation audit instruments provided in the Supplementary Materials. The table translates the protocol into a minimal set of actions students can complete when using AI-assisted literature review tools, with emphasis on prompt logging, DOI verification, metadata checking, duplicate control, and reproducibility review.

This compact checklist is intended to support classroom implementation by converting the audit protocol into a minimal verification routine that can be completed by students and reviewed by instructors.

2.7. Data Items

The audit database and supporting artifacts were prepared for public sharing under a Data Management and Sharing Plan (DMSP) (W. Wang et al., 2016). All records were

managed in a shared Zotero group and will be released as Supplementary Materials in CSV format at four decision points in the workflow: Dataset S3. Identification (All Runs, n = 212), Dataset S4. Screening Cleaned (n = 170), Dataset S5. Full-text eligibility (n = 100), and Dataset S6. Included Articles (n = 37). These staged exports preserve the audit trail from ingestion through inclusion and enable replication of deduplication, screening, and synthesis decisions. Materials will be deposited via the Open Science Framework, and derivative presentations will be disseminated through the institutional repository (Digital Commons STEM for Success). The following subsections define the outcome variables used for the audit and the additional variables required to reproduce analyses.

Table 1. Compact student-facing verification checklist for AI-assisted literature search workflows.

Step	Audit Question	Required Evidence
1	Was the tool, access tier, date, and full prompt recorded?	Prompt log with tool, mode, date, and browsing status.
2	Was the raw output preserved prior to any editing?	Fully saved output or export.
3	Does each record include a DOI or other stable identifier, or an explicit NO DOI notation?	DOI, PMID, stable URL, or NO DOI.
4	Does the DOI resolve, and does the landing page match the cited item?	Resolvable DOI and matching landing page.
5	Do the citation metadata match the source record?	Verified author(s), title, venue, and year.
6	Were duplicates removed before screening or synthesis?	Deduplicated library or CSV file.
7	Is the source eligible and in scope?	Peer-review status and inclusion decision recorded.
8	Was the prompt re-run to assess reproducibility?	Saved second run and comparison notes.
9	If drift occurred, was it documented?	Short explanation of likely cause.
10	Is there a complete audit trail for review?	Exportable files, notes, and completed checklist.

Note. The checklist is intended for instructional use and operationalizes the minimum verification actions required before AI-assisted outputs can be treated as usable review inputs.

2.7.1. Outcomes

The study measured three primary outcomes: (i) DOI traceability, coded using observable verification criteria; (ii) metadata integrity, operationalized as agreement between DOI landing-page metadata and citation fields (title, authors, venue, year); and (iii) run-to-run drift, operationalized as overlap metrics between paired runs.

2.7.2. Other Variables

Secondary variables included tool identity, run ID and pairing, retrieval posture (natural-language versus Boolean), yield per run, and bibliometric descriptors (Zhan et al., 2022) used for VOSviewer, version 1.8.0_202, profiling (co-authorship and keyword co-occurrence). Keyword provenance and normalization actions were explicitly coded using the codebook tags _KW-MISSING-PDF, _KW-VERIFIED-PUBLISHER, _KW-SUBSTITUTED-CATEGORIES, and _KW-CCS-CONCEPTS, with provenance notes retained in Zotero and cross-referenced in the screening log.

2.8. Risk of Bias Assessment (Tool-Output Bias)

The study assessed risk of bias as systematic integrity risk introduced by tool-produced bibliographic records, rather than bias within the underlying research literature. The

audit operationalized risk via observable failure modes, including DOI wrong-match, DOI non-resolving, bibliographic field instability (author, title, venue, year), inconsistent APA 7 author truncation conventions, and source-type substitution. Particular attention was given to author-name parsing errors that affect communities where two given names and/or two family names are commonly found, because these errors distort attribution, impair deduplication, and propagate downstream citation inaccuracies. The study treated these failure modes as workflow-level bias risks because they inflate perceived credibility while reducing verification and accountability.

2.9. DOI Verification Protocol (Traceability Audit)

The study verified DOI traceability through a fixed escalation procedure intended to maximize reproducibility and minimize discretionary searching. The team first opened the DOI link as provided. If the DOI did not resolve to the cited resource, the team searched by title; then by title plus journal or publisher; then by author; then by author plus title. If these steps did not resolve the record, the team navigated directly to the journal website and verified the article's presence using volume and issue. A record was coded DOI_Correct only when the DOI resolved, and the landing page matched title, authorship, venue, and year. Two cases required deeper inspection because the DOI and venue fields aligned, but the title differed materially; one remaining tie-break case was referred to the librarian for confirmation.

2.10. Effect Measures

The study quantified reproducibility and alignment using set-overlap measures. Run-to-run drift was computed using the Jaccard index over citation strings and, more conservatively, over DOI strings (Kannan et al., 2016). When run sizes differed, the study computed overlap/min as the intersection divided by the smaller set size to provide an interpretable turnover proxy. These measures supported tool-to-tool comparisons under the same top-*k* constraint.

2.11. Synthesis Methods

The study synthesized findings at three analytic levels. At the run level, the analysis compared completion and yield under fixed top-*k* capture constraints. At the record level, the analysis compared DOI traceability and metadata integrity failure-mode profiles across tools and retrieval postures. At the corpus level, the study profiled thematic structure using VOSviewer, version 1.8.0_202, applying a co-authorship threshold of at least three documents per author and a keyword co-occurrence threshold of at least five occurrences using full counting. CGT constant comparison and memo-writing supported category refinement across these levels by linking observed error patterns to workflow-relevant constructs (traceability, provenance, reproducibility, and accountable inquiry).

2.12. Reporting Bias Assessment

The study treated reporting bias as documentation and provenance bias introduced by tool mediation, including missing or weak rationales for query construction, opaque browsing and ranking processes, and substitution of non-peer-reviewed sources presented in journal-like formats. The assessment was narrative and anchored to observed failure modes and run logs, emphasizing how documentation gaps limit reproducibility even when outputs appear plausible.

2.13. Certainty Assessment

The study evaluated certainty for audit claims rather than effect estimates about the domain literature. Findings based on DOI-verified traceability and metadata integrity

were treated as high-confidence within the audited corpus because they relied on direct resolution checks and field-level matching. Conclusions about alignment with a librarian reference baseline were reported as bound to the current evidence and interpreted cautiously, because the librarian target set overlap analysis depends on an externally defined benchmark and can change with benchmark scope.

2.14. Use of Generative Artifitial Inteligence

Generative AI tools were used to generate candidate citations, propose Boolean queries, and draft explanatory rationales as part of the audited workflows (Shin, 2025). The research team performed manual verification, deduplication, screening decisions, and traceability coding using the described protocols. Grammarly was used for superficial language editing of prompts and manuscript syntax; this use did not alter study data, coding decisions, or analytic criteria. Figures and bibliometric maps were generated using Zotero exports and VOSviewer, version 1.8.0_202.

2.15. Data and Materials Availability

To enable replication and extension, the study will provide the canonical prompt, captured run logs and outputs, the cleaned Zotero library export and deduplicated CSV, the traceability codebook, and calculation sheets or scripts sufficient to reproduce the main descriptive statistics and overlap metrics to compare the clustering structures (Patriota et al., 2018).[3] Materials will be deposited in the Open Science Framework, with institutional-repository dissemination for presentation derivatives. Any temporary access restrictions will be disclosed at submission and removed prior to publication when feasible.

The Section 3 reports how the audited workflows performed across tools and retrieval postures, first by documenting record flow and dataset construction, then by quantifying DOI traceability, metadata integrity failure modes, and reproducibility metrics under the fixed top-*k* design.

3. Results

Section 3 reports the empirical outcomes of the audit and is organized into corpus construction and run completion (Sections 3.1 and 3.2), corpus characterization (Section 3.3), integrity diagnostics (Sections 3.4 and 3.5), and cross-run synthesis and reproducibility assessment (Sections 3.6–3.8). We begin by establishing a coherent, deduplicated corpus to support valid cross-tool comparisons.

3.1. Corpus Construction and Run Completion

Figure 1 summarizes the staged consolidation, screening, and inclusion workflow used to establish a common evidence base for cross-tool comparisons. After deduplication and LIS expert augmentation, the consolidated corpus comprised 197 records and was advanced to title and abstract screening, full-text assessment, and quality appraisal. Unless otherwise noted, descriptive profiling is based on the consolidated corpus ($n = 197$), whereas synthesis-level analyses are based on the included set ($n = 37$).

The authors maintained the consolidated library in a shared Zotero group to support team-based screening, adjudication, and metadata harmonization, and we plan to make the group publicly accessible upon publication to enable replication and reuse.

3.1.1. Run Completion and Captured-Item Yield

This subsection reports run completion and captured-item yield across the eight natural-language runs and their paired Boolean queries executed in Scopus.

Across natural-language runs, the tools produced 101 outputs, including one non-result placeholder (N/A). After excluding the non-result placeholder, the natural-language

posture yielded n = 100 captured items. Across Boolean executions, the tools produced 129 outputs, including one non-result placeholder (N/A). After excluding the Boolean non-result placeholder, Scopus execution yielded n = 128 retrievable items; Zotero then removed 8 duplicates during consolidation, yielding n = 120 unique Boolean-stage items. Combining both postures, the pre-consolidation corpus comprised 220 items (100 natural-language + 120 Boolean), which then entered the staged deduplication and screening workflow reported in Figure 1 and Section 2.5.

Table 2 summarizes captured-item yield by run and indicates whether the k = 20 target was met. When a Boolean query returned zero results, the query was shortened by reducing the final clause segment first while preserving the core construct families; all such revisions are documented and red-coded in the Supplementary Materials to support reproducibility.

Table 2. Retrieval yield by run (natural-language vs. Boolean).

Run	Tool	Type	Retrieved (n)	Notes
Run 1	Gemini (institutional)	Natural-language	18	Complete list produced
Run 1.1	Gemini (institutional)	Boolean	20	Target met
Run 2	Gemini (institutional)	Natural-language	18	Complete list produced
Run 2.1	Gemini (institutional)	Boolean	20	Target met
Run 3	ChatGPT free tier	Natural-language	8	Under-target yield
Run 3.1	ChatGPT free tier	Boolean	20	Target met
Run 4	ChatGPT free tier	Natural-language	5	Under-target yield
Run 4.1	ChatGPT (Free tier)	Boolean	N/A [1]	No retrievable items captured (logged)
Run 5	Perplexity (Academic configuration)	Natural-language	N/A [1]	No retrievable items captured (logged)
Run 5.1	Perplexity (Academic configuration)	Boolean	15	Complete list produced
Run 6	Perplexity (Academic configuration)	Natural-language	7	Under-target yield
Run 6.1	Perplexity (Academic configuration)	Boolean	13	Complete list produced
Run 7	ChatGPT paid tier (Extended Thinking)	Natural-language	19	Complete list produced
Run 7.1	ChatGPT paid tier (Extended Thinking)	Boolean	20	Target met
Run 8	ChatGPT paid tier (Extended Thinking)	Natural-language	17	Complete list produced
Run 8.1	ChatGPT paid tier (Extended Thinking)	Boolean	20	Target met

Note. "Target met" denotes k = 20 captured items. Deduplication is reported separately in the corpus flow (see Figure 1). [1] "non-result placeholder (N/A), excluded from item counts." Natural-language captured items total (excluding N/A): 100; Natural-language after Zotero deduplication: 92; Boolean retrievable items total (excluding N/A): 128.

Across tools, Boolean translation increased yield consistency relative to natural-language prompting, including for tools that under-produced in natural-language mode. The only runs with zero captured items were ChatGPT free tier, Boolean Run 4.1, and Perplexity (Academic configuration) Natural-language Run 5. In the ChatGPT free tier natural-language condition (Runs 3 and 4), the system explicitly indicated that it could not fully complete the request as specified and instead offered a smaller set of "possible sources." This self-reported limitation is consistent with the under-target yields observed in those runs. Having established run completion and bounded yields, we next evaluate DOI-based traceability to determine which retrieved citations support verifiable, accountable sourcing.

3.1.2. DOI Traceability Outcomes

Traceability was operationalized as a two-step criterion: (a) whether a DOI string was provided and (b) when a DOI was provided, whether it both resolved and matched the cited bibliographic record (title, authors, venue, year). For natural-language outputs, traceability failures clustered into three recurring patterns: missing DOIs (NO_DOI), non-resolving or malformed DOIs (DOI_NON_RESOLVING), and resolving DOIs that corresponded to a different work than the accompanying citation metadata (DOI_WRONG_MATCH). Because these failure types are observable and reproducible, we report them as a codebook with explicit operational definitions (see Table 3), which supports transparent auditing and anticipates critiques that AI-related claims are impressionistic rather than evidence-based.

Table 3. Traceability failure codebook (operational definitions).

Code	Operational Definition	Observable Indicator in Audit
DOI_Correct	DOI resolves and matches title, authors, venue, and year	DOI landing page aligns with citation fields
DOI_NON_RESOLVING	DOI does not resolve (or is malformed)	DOI returns an error, redirects incorrectly, or is not a DOI
DOI_WRONG_MATCH	DOI resolves, but metadata does not match citation	DOI resolves to a different paper, journal, or year
NO_DOI	No DOI provided, or source type lacks DOI	Conference volumes, blogs, technical reports, or incomplete records

Note. Personal elaboration.

We then applied these codes to the natural-language and Boolean outputs (see Table 4).

Natural-language retrieval produced substantial traceability degradation in Gemini (institutional) and Perplexity (Academic configuration), primarily through non-resolving and wrong-match DOIs, which create a high-risk "false verification" condition (a DOI appears present but does not verify the cited work). Under the same DOI-required criterion, ChatGPT 5.2 (Extended Thinking) produced markedly higher DOI-correct rates, indicating greater stability in DOI provision and bibliographic field alignment.

When the same topical intent was translated into a Scopus-format Boolean query and executed as database retrieval, the resulting records exhibited higher DOI availability and higher DOI-correct rates (see Table 5). Residual NO_DOI outcomes reflect records that lack DOIs in the database export or belong to item types where DOI assignment is inconsistent.

Table 4. Traceability outcomes by tool (natural-language runs only).

Tool	Retrieved (n)	DOI_Correct (n)	DOI_Correct (%)	Dominant Failure Modes
Gemini (institutional)	40	9	22.5%	DOI_NON_RESOLVING, DOI_WRONG_MATCH, NO_DOI
ChatGPT free tier	13	9	69.2%	NO_DOI (not errors, but non-DOI sources)
Perplexity (Academic configuration)	7	2	28.6%	DOI_NON_RESOLVING, DOI_WRONG_MATCH
ChatGPT paid tier (Extended Thinking)	40	38	95.0%	Rare DOI mismatch and resolvability failure

Note. DOI_Correct indicates that the DOI resolved and matched the citation metadata; DOI_WRONG_MATCH indicates a resolving DOI associated with a different record; DOI_NON_RESOLVING indicates a DOI string that failed to resolve; NO_DOI indicates that no DOI string was provided.

Table 5. Traceability outcomes by tool (boolean runs only).

Tool	Retrieved (n)	DOI_Correct (n)	DOI_Correct (%)	Dominant Failure Modes
Gemini (institutional)	40	32	80.0%	NO_DOI
ChatGPT free tier	21	20	95.2%	Run 4.1 generated 0 results
Perplexity (Academic configuration)	28	23	82.1%	NO_DOI
ChatGPT paid tier (Extended Thinking)	40	40	100.0%	No DOI mismatch and no resolvability failures

Note. These counts summarize DOI fields for records retrieved via Boolean queries. DOI_Correct (%) is computed as DOI_Correct divided by Retrieved; NO_DOI denotes absent DOI fields in the exported result set.

Having established run-level yield and DOI traceability as baseline integrity indicators, we next classify excluded records to make the eligibility boundary of the consolidated corpus explicit before characterizing themes and synthesis patterns.

3.1.3. LIS Expert Baseline for Topical Scoping and Exclusion Boundaries

This subsection defines the librarian-derived reference set used to operationalize the study's search topic and adjudicate ambiguous inclusion decisions. An LIS expert executed Boolean searches (see Appendix A.1) in Scopus, Web of Science, and Google Scholar and applied database-specific limits on publication type, date, and language when available (Table 6). Records were screened on title and abstract, and then assessed at full text for topical fit. Studies were excluded if they did not connect AI, NLP, or LLM-enabled tools to literature searching practices (query formulation, retrieval, or citation handling) within evidence synthesis or review contexts, or if the paper treated AI systems as the object of review without examining their role in the search workflow. In the full text, additional exclusions removed studies that did not examine AI-based tools used for the literature search stage.

Table 6. Librarian baseline search parameters and exclusion criteria by database.

Database	Boolean Search	Limits/Filters	Exclusion Criteria
Scopus	(AI OR artificial intelligence OR LLM OR large language model) AND (literature review OR systematic review OR literature survey) in ARTICLE TITLE, ABSTRACT, KEYWORDS	Document type: article; Language: English; Date: 2020 to 2026	AI is the subject of the literature review rather than being utilized as a tool; the article focuses on AI for writing; AI tools are not used in the literature search
Web of Science	(AI OR artificial intelligence OR LLM OR large language model) AND (literature review OR systematic review OR literature survey) in all databases	Document type: article; Language: English; Date: 2020 to 2026	AI is the subject of the literature review rather than being utilized as a tool; AI tools are not used in the literature search
Google Scholar	(AI OR artificial intelligence OR LLM OR large language model) AND (literature review OR systematic review OR literature survey)	Date range: since 2023; results sorted by relevance	AI is the subject of the literature review rather than being utilized as a tool; AI tools are not used in the literature search

Note. The baseline operationalizes the study's search-stage topic.

The LIS expert source set (Dataset S7; n = 27) served as an external baseline to ground thematic scope and operationalize exclusion decisions during screening. We profiled Dataset S7 to identify recurrent constructs, canonical terminology, and foundational authors in the literature on AI-assisted searching, systematic review automation, librarian benchmarks, and evidence synthesis methods. This baseline specified the core construct families used to delimit relevance (for example, large language models, natural language processing, systematic literature review, artificial intelligence, ChatGPT, and literature review) and supported differentiation between central workflow-focused studies and peripheral or domain-specific applications.

We used the expert baseline in two ways. First, it supported negative case identification by specifying exclusion cues, including records focused on AI applications unrelated to literature review workflows, sources lacking methodological detail relevant to traceability and reproducibility, and studies centered on educational uses of AI without a review-workflow component. Second, it provided an author and citation structure for assessing topical alignment in ambiguous cases or disagreement, facilitating iterative team discussions (Göndöcs et al., 2025). Because co-authorship patterns in small expert sets can be contingent on publication teams, we treated co-authorship mapping as descriptive context. We relied on co-citation and reference frequency within Dataset S7 to identify high-salience methodological anchors (for example, systematic review automation and living review methods). These anchors informed adjudication when tool-derived records presented ambiguous relevance. The resulting exclusion boundary was applied during title and abstract screening and full-text assessment, with all decisions logged in the screening log and traceability codebook. Supplementary Materials include Figures S1 and S2, which present baseline, expert-seeded diagnostic VOSviewer, version 1.8.0_202, overlays (keyword co-occurrence and co-authorship) for Dataset S7.

3.2. Exclusion Logic, Screening Outcomes, and Metadata Completion

This study treated tool outputs as the object of analysis and therefore retained all citations returned by the AI-assisted tools at the point of capture. Exclusions were not applied as a relevance filter at this stage; instead, records were later categorized by trace-

ability status and deduplicated for corpus consolidation. The only immediate exclusions in the natural-language condition were cases where a tool-generated citation could not be mapped to any identifiable bibliographic record during manual verification. Specifically, four Gemini (institutional) citations (two per run) yielded no identifiable record. In ChatGPT 5.2 (Extended Thinking) natural-language outputs, one citation in Run 7 and three citations in Run 8 could not be verified to an identifiable record under the DOI-required rule. These cases were retained in the dataset as untraceable items and coded accordingly rather than removed.

For Boolean retrieval, exclusions were operationalized as query-level feasibility constraints rather than study-level relevance decisions. When possible, the whole Scopus-format Boolean string was executed to retrieve $k = 20$ records. In Perplexity (Academic configuration) Boolean Runs 5.1 and 6.1, the whole query string returned zero results; therefore, the Boolean string was shortened to restore retrievability. For Boolean runs that produced result sets larger than $k = 20$, the exported lists were filtered by citation-based ordering (highest cited first), and only the top 20 records were retained to preserve cross-run comparability.

3.2.1. Eligibility Screening Criteria and Exclusion Coding

Eligibility was operationalized using the inclusion and exclusion criteria defined in Section 2.2 and applied consistently to all records in the consolidated corpus, including the 27 records contributed by the LIS expert. Screening proceeded in two stages: (I) Titles and abstracts were screened to remove clearly ineligible records. (II) Full texts were assessed to confirm topical relevance, methodological fit with the audit focus, and sufficient reporting detail to support traceability assessment.

Exclusion codes were developed iteratively using the expert baseline (Dataset S7) and finalized in the codebook. All screening decisions and rationales were recorded in the screening log and traceability codebook.

3.2.2. Title and Abstract Exclusions

Records were excluded at the title and abstract stage for the following primary reasons (mutually exclusive reason codes; see screening log for definitions and examples):

- Out of scope topic (AI not applied to literature review or evidence synthesis workflows) (n = [T1]).
- Non-empirical or non-methodological commentary without evaluable workflow content (n = [T2]).
- Publication type ineligible for this review (for example, editorial, news item, non-scholarly blog content when peer-reviewed evidence was required) (n = [T3]).
- Not focused on traceability, metadata integrity, reproducibility, or comparable workflow quality constructs (n = [T4]).
- Duplicate or near duplicate not captured by automated deduplication (n = [T5]).
- Other pre-specified exclusion reason (n = [T6]).

3.2.3. Full Text Exclusions and Quality Appraisal Removals

Full-text assessment was conducted for 100 records. Exclusions at this stage reflect ineligibility confirmed through full-text review and removals after quality appraisal. Primary exclusion categories were:

- Full text confirmed out of scope, despite apparent relevance in abstract (n = [F1]).
- Insufficient methodological transparency to support audit and traceability assessment (for example, missing or irreproducible search description, undefined selection process, absent tool configuration details) (n = [F2]).

- Ineligible study type or evidentiary basis under the protocol (n = [F3]).
- Non-retrievable full text or unusable reporting format (n = [F4]).
- Quality appraisal threshold not met (n = [F5]).
- Other pre-specified full-text exclusion (n = [F6]).

3.2.4. Adjudication of Ambiguous Cases

Ambiguous relevance decisions were resolved through adjudication using the screening log and codebook. When initial judgments diverged, reviewers compared the record against the inclusion criteria, identified the specific rule in contention, and recorded the rationale for the final decision. Complex cases were escalated to a senior reviewer for a tie break, and the adjudication outcome was logged to preserve decision provenance.

3.2.5. Traceability-Related Metadata Completion

To improve metadata integrity and reduce downstream ambiguity in bibliometric profiling, records were normalized in Zotero prior to analysis. Where author-provided keywords were missing from the PDF, keywords were verified from the journal website and recorded in Zotero with provenance noted (n = 5 cases). For the single non-journal web-based item retained for contextual coverage, platform categories were recorded instead of author keywords with provenance documented (n = 1 case).

3.3. Corpus Characteristics and Descriptive Bibliometric Profile

To characterize the consolidated corpus beyond run-level integrity counts, we used VOSviewer, version 1.8.0_202, as a descriptive bibliometric profiling tool (Bukar et al., 2023). VOSviewer, version 1.8.0_202, supports network mapping of bibliographic relations (for example, co-authorship ties) and semantic relations (for example, keyword co-occurrence), enabling a compact assessment of whether the corpus is structurally concentrated or thematically fragmented (Ballesteros-Ballesteros & Zárate-Torres, 2025). We produced two complementary outputs: (a) a co-authorship overlay map to examine the concentration of productivity and collaboration within a small author cluster, and (b) a keyword co-occurrence overlay map to identify recurrent constructs and their temporal distribution.

3.3.1. Co-Authorship Structure

We used VOSviewer, version 1.8.0_202, as a descriptive bibliometric tool to characterize the cleaned corpus and assess whether the evidence base was dominated by a small set of authors. A co-authorship analysis was conducted using a minimum threshold of three documents per author (see Figure 2). Of 721 authors, 8 met this threshold, indicating that the corpus was not dominated by a single author-centric community, but instead spanned multiple adjacent literature bases.

The co-authorship network was generated in VOSviewer, version 1.8.0_202, using the cleaned corpus (n = 190) with a minimum threshold of three documents per author. Of 721 authors, 8 met the threshold and are displayed. Node size reflects author document count; links indicate co-authorship ties; and link thickness reflects collaboration strength. Node color reflects the average publication year (overlay scale), illustrating the temporal distribution of the most prolific authors in the corpus. The sparse network indicates that a single tightly connected author community does not dominate the dataset, but rather spans multiple adjacent strands of evidence synthesis, information retrieval, and AI-enabled workflow research.

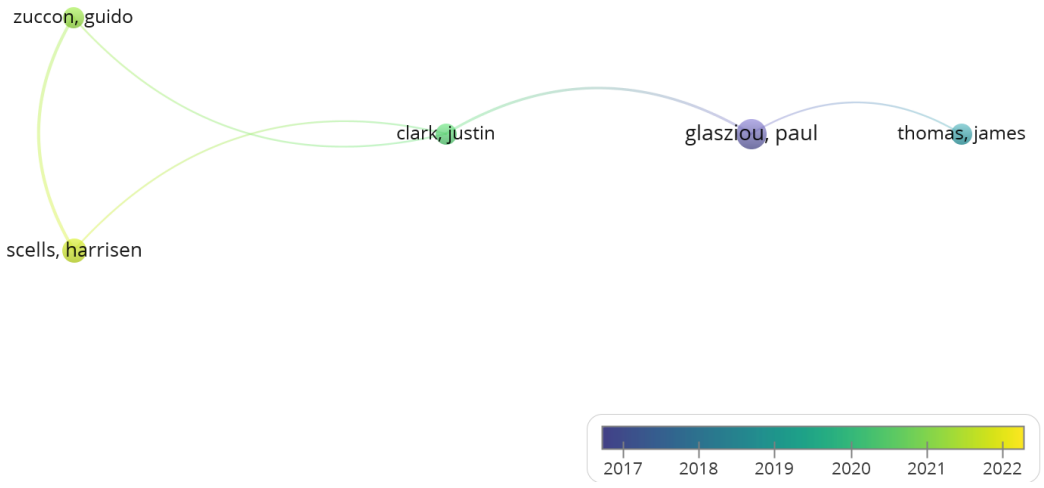

Figure 2. Co-authorship analysis map (VOSviewer, version 1.8.0_202, overlay visualization).

3.3.2. Keyword Co-Occurrence Themes and Temporal Structure

A keyword co-occurrence analysis was conducted using "full counting" with a minimum keyword occurrence threshold of five (Figure 3). Of 561 keywords, 13 met this threshold, indicating a compact set of recurrent constructs adequate for describing the corpus while avoiding over-interpretation of thematic coherence.

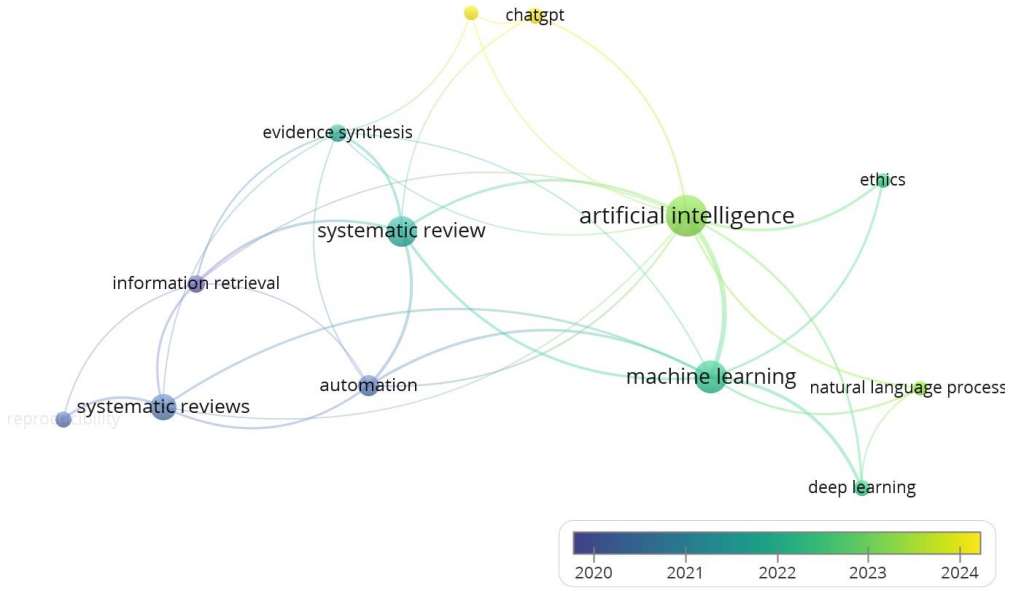

Figure 3. Keyword co-occurrence analysis map (VOSviewer, version 1.8.0_202, overlay visualization).

The keyword co-occurrence network was generated in VOSviewer, version 1.8.0_202, from the consolidated corpus (n = 190) using full counting and a minimum threshold of five occurrences per keyword. Of 561 keywords, 13 met the threshold and are displayed. Node size reflects keyword frequency; links indicate co-occurrence within the duplicate records; and link thickness reflects co-occurrence strength. Node color represents the average publication year (overlay scale), enabling a temporal reading of thematic emphasis. The map indicates an early concentration (circa 2020–2021) around systematic review(s) and information retrieval, followed by increased salience of machine learning (circa 2022) and a shift toward artificial intelligence and natural language processing (circa 2023), with ChatGPT emerging as a named focal term in the most recent period (circa 2024).

The overlay progression suggests a cumulative convergence rather than a topic replacement. Early work centers on methodological infrastructure (systematic review procedures and information retrieval). Later work introduces automation and machine learning as enabling techniques, which then expands into broader "artificial intelligence" framing and, most recently, tool-specific discourse (ChatGPT) that foregrounds the implications of generative models for retrieval, synthesis, and citation practices. This temporal structure supports interpreting the corpus as an evolving methods-and-tools conversation about accountable evidence work, rather than as a purely AI-in-education outcomes literature.

3.4. Integrity Risks in Tool-Produced Bibliographic Records

This subsection evaluates threats to validity arising from AI-assisted generation of bibliographic records under natural-language prompting. Because the unit of analysis is the tool-produced citation (not only the underlying publication), bias in this study primarily manifests as systematic distortion in bibliographic fields, identifier assignment, and source-type labeling. These distortions can inflate apparent rigor (e.g., "complete-looking" citations) while reducing auditability, thereby biasing any downstream comparison of relevance, overlap, or synthesis quality.

Bias Risks in Tool-Produced Bibliographic Records

Across tools, natural-language prompting produced recurring metadata integrity failures that materially weaken accountable inquiry, even when citations superficially appeared "complete." The most consequential error class was DOI wrong-match, in which a DOI resolved but pointed to a different work than the cited title, authorship, venue, or year. This failure is high-risk because it creates a false verification signal and can pass casual checks. A second frequent error class was DOI non-resolving, indicating malformed DOI strings, transcription errors, or incorrect DOI assignment. A third pattern was bibliographic field instability, including author and title misattribution, incorrect venue attribution, and inconsistent application of APA 7 conventions (for example, misuse of "et al.", inconsistent author ordering, or inconsistent truncation). A fourth pattern was source-type substitution, where non-peer-reviewed web or organizational content was presented in journal-like form, inflating perceived scholarly legitimacy while undermining traceability requirements.

In addition, several tools showed systematic fragility when handling author names that include compound given names and/or compound surnames (e.g., two surnames, hyphenation, particles, or culturally specific ordering). These cases increased the frequency of author-field corruption and contributed to downstream mismatches during manual verification. More broadly, most tools produced citations that deviated from APA 7 formatting requirements at a non-trivial rate; in the audited runs, ChatGPT 5.2 (Extended Thinking) exhibited the highest overall accuracy in citation formatting and field stability relative to the other tools.

These patterns demonstrate why DOI presence is not equivalent to metadata integrity and why integrity must be evaluated as a separate construct from retrieval yield. All returned citations were manually entered into Zotero and cross-checked against DOI landing pages or authoritative source records to identify mismatches and to assign traceability failure codes. Under these verification conditions, ChatGPT 5.2 (Extended Thinking) exhibited the most substantial alignment between tool-produced citations and human verification. At the same time, Gemini (institutional) and Perplexity (Academic configuration) produced higher rates of DOI resolvability and matching errors, consistent with their higher frequencies of DOI_NON_RESOLVING and DOI_WRONG_MATCH in the run-level audit.

Having specified the principal bias risks that affect interpretability, we next report run-level results for each tool and condition, including metadata integrity profiles, failure-mode distributions, and reproducibility indicators.

3.5. Run-Level Integrity and Traceability Results

This subsection reports run-level metadata integrity and DOI-traceability outcomes. We focus on two integrity dimensions that condition accountable inquiry: (a) metadata integrity, operationalized through traceability failure modes indicating where bibliographic fields and identifiers break down, and (b) DOI-based traceability, operationalized as the proportion of returned items that can be verified through a resolving DOI whose landing-page metadata matches the cited record. Reporting both dimensions is necessary because a citation can appear syntactically well-formed while still failing verification (for example, DOI wrong-match), and because yield alone does not indicate whether retrieved items can support reproducible evidence work. Supplementary Table S1 provides the complete run log and captured outputs, organized by natural-language runs, Boolean executions, and the combined item-level audit table.

3.5.1. Librarian Baseline Search Yields

The expert librarian search produced distinct but complementary result sets across Scopus, Web of Science, and Google Scholar. In Scopus, the Boolean query limited to article title, abstract, and keywords retrieved 247 records. Filtering by document type reduced this to 93 articles, language limits to English reduced it to 89, and the date range 2020 to 2026 reduced it further to 87 candidate records.

In Web of Science, the same Boolean query executed across all databases returned 89 records. Refinement by publication year (2020 to 2026), document type (article), and language (English) reduced the set to 43 records. From this refined set, an initial subset of 3 records was pulled for deeper analysis, followed by an additional 6 records flagged in a subsequent pass for closer review.

In Google Scholar, the query initially returned approximately 51,000 results. Applying a custom date range since 2020 reduced this to 19,400 results, which were then sorted by relevance. From this ranked list, 20 records were pulled for further analysis.

All records selected for deeper review were added to Zotero and screened using a consistent inclusion criterion: AI needed to be used in literature searching, not only in writing, data extraction, synthesis, or any other component of the review process. From this corpus, the final librarian expert set, Dataset S7 Expert Sources, was compiled with 27 sources (n = 27).

3.5.2. Metadata Integrity and Failure-Mode Profiles

Table 7 decomposes natural-language performance at the run level, showing that traceability failures cluster by tool family rather than occurring uniformly across runs. Together, the run-level counts and the failure-mode definitions provide the empirical basis for interpreting tool differences as measurable, reproducible integrity outcomes rather than anecdotal impressions of "AI quality."

Natural-language traceability outcomes were highly heterogeneous across runs. Practically, this means that students using conversational retrieval cannot assume that a 'complete-looking' citation is verifiable without manual checking. Gemini (institutional) runs exhibited large proportions of DOI_NON_RESOLVING and DOI_WRONG_MATCH failures, indicating unstable DOI assignment and bibliographic mismatches. Perplexity (Academic configuration) displayed one complete zero-yield run and, when results were produced, a mixed failure profile with both non-resolving and wrong-match DOIs. By contrast, ChatGPT free tier runs under-produced items but yields comparatively higher traceability

among returned items. At the same time, ChatGPT 5.2 (Extended Thinking) produced both the highest yields and the highest traceability rates, culminating in a fully traceable run (Run 8) under manual verification. Aggregated across all natural-language runs (n = 100 items), only 58% met the DOI_Correct criterion, with the remainder distributed across non-resolving DOI, wrong-match DOI, and no-DOI outcomes. This distribution indicates that integrity failures are not marginal edge cases; they constitute a structurally meaningful portion of tool-produced outputs in natural-language retrieval.

Table 7. Tool natural-language traceability failure modes (DOI required).

Run	DOI Correct	DOI NON RESOLVING	DOI WRONG MATCH	NO DOI	Traceability Rate (%)
Natural-language Run 1	1	8	6	5	5%
Natural-language Run 2	8	7	4	1	40%
Natural-language Run 3	5	0	0	3	63%
Natural-language Run 4	4	0	0	1	80%
Natural-language Run 5	0	0	0	0	0%
Natural-language Run 6	2	3	2	0	29%
Natural-language Run 7	22	1	1	0	92%
Natural-language Run 8	16	0	0	0	100%
Total (n = 100)	58	19	13	10	58%

Note. Run-level distribution of DOI-based traceability outcomes across tools under natural-language prompting. DOI_Correct indicates a resolving DOI whose landing-page metadata matches the cited title, authors, venue, and year. DOI_NON_RESOLVING indicates a malformed or non-resolving DOI string. DOI_WRONG_MATCH indicates a resolving DOI that corresponds to a different work than the cited metadata. NO_DOI indicates no DOI provided or a source type that does not supply a DOI. Traceability rate is the percentage of items in the run coded DOI_Correct.

3.5.3. Traceability Outcomes by Run

Interpreted at the run level, the traceability distribution reveals two practically relevant patterns. First, some runs failed primarily through identifier instability (DOI_NON_RESOLVING), suggesting that the tools generated plausible DOI strings that could not be verified. Second, other runs failed through false verification (DOI_WRONG_MATCH), a higher-risk mechanism in which a DOI resolves successfully but validates a different publication than the cited record. From an accountability perspective, DOI_WRONG_MATCH failures are more damaging than NO_DOI because they can survive cursory screening and propagate into downstream synthesis as "verified" evidence. Conversely, runs with high DOI_Correct proportions indicate conditions where tool outputs can be meaningfully audited and reproduced, enabling stronger claims about workflow traceability.

3.6. Cross-Run Synthesis and Reproducibility Results

This subsection synthesizes results across retrieval postures and runs to compare traceability, metadata integrity, and reproducibility.

3.6.1. Traceability Performance by Retrieval Posture

Table 8 reports run completion and aggregate traceability performance by retrieval posture, showing that Boolean translation increased both yield and DOI-traceability relative to natural-language prompting.

As shown in Table 7, the design specified up to 160 items per condition (8 runs $\times$ k = 20). Natural-language prompting yielded 101 items (63.1% of the maximum), reflecting multiple under-target runs and one complete zero-yield event. Boolean translation and retrieval yielded 129 items (80.6% of the maximum), indicating improved completion and better compliance with the top-k capture requirement. This improvement is not only quantitative. Boolean-mode retrieval also shifted the corpus toward DOI-bearing records, thereby increasing verifiability under a DOI-required traceability standard (subject to subsequent

resolution and metadata matching checks). In synthesis terms, Boolean translation functioned as a governance mechanism: it reduced reliance on implicit ranking heuristics and increased the proportion of outputs that can be independently audited.

Table 8. Tool run inventory and completion.

Condition	Runs	k per Run	Max Items	Total Items	Correct (%)	Notes
Natural-language prompting	8	20	160	92	57.5%	DOI presence assessed from Tool output; traceability assessed post-manual verification
Boolean translation and retrieval	8	20	160	128	80.0%	Scopus-format query + rationale; DOI-traceability coded from result lists

Note. Run completion and item yield by retrieval posture. "Max items" indicates the theoretical maximum (8 runs × 20). "Correct (%)" reports the proportion of the maximum retrieved under each posture. In the natural-language condition, DOI presence was recorded as printed by the tool output, and traceability was determined after manual verification. In the Boolean condition, Scopus-format queries and their rationales were captured, and DOI-traceability was coded from the resulting item lists using the DOI-required rubric.

3.6.2. Overlap with Librarian Reference Set

This subsection estimates alignment between AI-retrieved outputs and a librarian benchmark using DOI-level identity. The librarian benchmark was derived from the LIS expert baseline (Dataset S7; $n = 27$; Supplementary Dataset S1) and operationalized as a DOI-verified target set BBB comprising the 26 items in Dataset S7 with resolvable DOIs that passed the study's DOI correctness checks ($|B| = 26 | B | = 26 |B| = 26$). For each AI run (natural-language and Boolean runs per tool), overlap was computed against this fixed target using DOI-level matching and was bounded to items coded DOI_Correct to avoid rewarding non-traceable outputs (see Table 9).

We report (i) overlap count, defined as the number of DOI identities shared between an AI run and the librarian target set; (ii) Jaccard similarity; and (iii) the overlap coefficient (Franco-Pereira et al., 2021). Let AAA denote the set of DOI_Correct items retrieved in a given run (bounded to top $k = 20$, $k = 20$, $k = 20$ when available), and let BBB denote the librarian DOI-verified target set. Jaccard similarity is defined as (Suanet, 2017, p. 5):

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

We also use the overlap coefficient (also called the Szymkiewicz–Simpson coefficient):[4]

$$overlap(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \quad (2)$$

We additionally report bounded precision and bounded recall at $k = 20$, $k = 20$, $k = 20$, where bounded precision is computed over DOI_Correct items in AAA and bounded recall is computed over DOI-verified items in BBB. Finally, we characterize mismatches qualitatively as systematic omissions (target DOIs never retrieved) versus systematic false positives (retrieved DOIs outside the target set), with adjudication rules and reason codes recorded in the screening log and traceability codebook (Supplementary Material: Document S4).

Table 9. Overlap with librarian target set (natural-language runs).

Run	A_Size_DC Correct	Overlap_ Count	Jaccard_ Overlap	Overlap_ Coefficient	Notes
1	18	0	0	0	No DOI-level match vs. librarian DOIs
2	18	0	0	0	No DOI-level match vs. librarian DOIs
3	8	1	0.030303	0.125	Match: (Bolaños et al., 2024)
4	5	0	0	0	No DOI-level match vs. librarian DOIs
5	0	0	0	N/A	Non-result placeholder; no retrievable items captured
6	7	1	0.03125	0.142857	Match: (Wu et al., 2025)
7	19	0	0	0	No DOI-level match vs. librarian DOIs
8	17	0	0	0	No DOI-level match vs. librarian DOIs

Note. DOI-bounded; |B| = 26 (librarian DOI-verified target set). |A| is the count of DOI_Correct items captured in each run (bounded to top k = 20 when available). Jaccard overlap is computed using Equation (1). Overlap coefficient is computed using Equation (2). For Run 5, the overlap coefficient is not defined because |A| = 0.

The Jaccard similarity measure is comparable to simple matching similarity, but ignores nonoccurrence frequency in the calculation ADDIN ZOTERO_ITEM CSL_CITATION (Kannan et al., 2016; Kotu & Deshpande, 2019). For the illustrative example X = (1,1,0,0,1,1,0) and Y = (1,0,0,1,1,0,0), the Jaccard index is computed using Equation (1).

3.6.3. Run-to-Run Drift

To estimate reproducibility under repeated execution, we operationalized run-to-run drift as the degree of overlap between paired runs conducted with the same tool under the same retrieval posture. Overlap was computed using the Jaccard index[5] ($|A \cap B| / |A \cup B|$). Because citation strings can differ by formatting, truncation, or APA-style variation, we report two overlap measures for natural-language runs: (a) citation-string overlap and (b) DOI-string overlap, treating DOI overlap as the more conservative indicator of stability (Tan et al., 2006). We also report overlap/min, defined as $|A \cap B|$ divided by the smaller of the two sets (Abdullah, 2016), to provide an interpretable turnover proxy when run sizes differ (Tables 8 and 9).

Table 10 indicates substantial heterogeneity in reproducibility across tools under natural-language prompting. Gemini (institutional) exhibited near-complete turnover (citation Jaccard = 0.00; DOI Jaccard = 0.057), consistent with high instability in the identifier-level composition of the top-k set. ChatGPT 5.2 (Extended Thinking) also showed low stability between runs (citation and DOI Jaccard = 0.059), despite high traceability rates within individual runs, demonstrating that per-item correctness does not imply run-to-run reproducibility of the retrieved set. ChatGPT free tier displayed comparatively high overlap across under-target runs (DOI Jaccard = 0.80; overlap/min = 1.00), which likely reflects a constrained output space rather than superior retrieval reproducibility; with short lists, overlap can be inflated. Perplexity (Academic configuration) produced a complete zero-yield run, which forces drift to 0 by definition and should be interpreted as a run-completion failure rather than as a meaningful stability.

Table 10. Run-to-run drift for natural-language outputs (paired runs within each tool).

Tool	Run A	nA	Run B	nB	Citation Overlap (n)	Citation Jaccard	DOI Overlap (n)	DOI Jaccard	DOI Overlap/min
Gemini (institutional)	Run 1	20	Run 2	20	0	0.0	2	0.057	0.118
ChatGPT free tier	Run 3	8	Run 4	5	5	0.625	4	0.8	1.0
Perplexity (Academic configuration)	Run 5	0	Run 6	7	0	0.0	0	0.0	
ChatGPT paid tier (Extended Thinking)	Run 7	20	Run 8	16	2	0.059	2	0.059	0.125

Note. Jaccard is computed as $|A \cap B| / |A \cup B|$. Overlap/min is computed as $|A \cap B|$ divided by the smaller set size. Rows with a zero-yield run should be interpreted as run-completion failures rather than stability estimates.

Table 11 shows that Boolean translation generally improved reproducibility relative to natural-language prompting within tools, though stability remained limited. Gemini (institutional)'s Boolean runs achieved moderate identifier stability (DOI Jaccard = 0.333), indicating a larger stable core under Boolean constraints. Perplexity (Academic configuration)'s Boolean runs showed similar moderate stability (DOI Jaccard = 0.278). ChatGPT 5.2 (Extended Thinking) exhibited lower stability in Boolean mode (DOI Jaccard = 0.111), suggesting that Boolean constraints alone do not guarantee stable retrieval composition when ranking and selection remain sensitive to execution conditions. For ChatGPT free tier, the second Boolean run was not available in the captured dataset ($n = 0$), which precludes drift estimation and is reported as a missing-run condition rather than as evidence of instability.

Table 11. Run-to-run drift for Boolean retrieval results (paired runs within each tool).

Tool	Run A	nA	Run B	nB	DOI Overlap (n)	DOI Jaccard	DOI Overlap/min
Gemini (institutional)	Run 1.1	20	Run 2.1	20	8	0.333	0.533
ChatGPT free tier	Run 3.1	20	Run 4.1	0	0	0.0	
Perplexity (Academic configuration)	Run 5.1	15	Run 6.1	13	5	0.278	0.455
ChatGPT paid tier (Extended Thinking)	Run 7.1	20	Run 8.1	20	4	0.111	0.2

Note. Drift was computed using DOI overlap. Jaccard is computed as $|A \cap B| / |A \cup B|$. Overlap/min is computed as $|A \cap B|$ divided by the smaller set size. For ChatGPT free tier, the second Boolean run was not available in the captured dataset ($n = 0$), which precludes drift estimation and is treated as missingness rather than instability.

3.6.4. Boolean Query Construct Coverage and Rationale Quality

Across tools, Boolean translation generally produced Scopus-format queries accompanied by rationales that articulated term inclusion and scope constraints. With one exception (ChatGPT free tier Run 4.1 was not captured), Boolean runs yielded analyzable queries and result lists. At the construct level, queries showed strong alignment with the locked research question by consistently including: (a) AI system terms (e.g., artificial intelligence, LLM, ChatGPT), (b) review-workflow terms (literature review, systematic review, evidence synthesis), (c) integrity and reproducibility terms (traceability, metadata, reproducibility, provenance, rigor), and (d) student population terms (undergraduate, graduate, student). These four construct families operationalize the minimum retrieval specification for auditing undergraduate and graduate literature review workflows.

Construct fidelity, however, was not uniformly complete. The most consequential omission observed in at least one Boolean run was the baseline comparator construct (librarian, information literacy, manual search), which is required to operationalize the "compared to a librarian reference baseline" clause of the research question. When baseline terms were absent, the query remained syntactically valid but became conceptually under-specified for the intended comparison, increasing the probability that retrieved records would skew toward generic AI-in-education outcomes or general systematic review automation rather than benchmarked workflow governance.

Rationale quality also varied in ways that matter for auditability. Higher-quality rationales explicitly (i) justified construct families by linking them to the study's outcome variables (traceability, metadata integrity, reproducibility), (ii) stated exclusion intent (e.g., avoiding generic AI ethics papers not tied to retrieval workflows), and (iii) acknowledged precision–recall trade-offs. Lower-quality rationales tended to restate the query without documenting why terms were selected or how the query enacts the librarian baseline and student context. Accordingly, we treat the rationale as a reproducibility artifact: it functions as an explicit warrant for construct inclusion/exclusion and should be evaluated as part of the protocol, not as narrative filler.

3.6.5. Implication for the Audit Design

Boolean translation improves completion and increases the proportion of DOI-bearing outputs, but it does not automatically ensure fidelity to the comparator structure of the research question. A minimal Boolean quality checklist should therefore include: (1) Scopus field tagging (TITLE-ABS-KEY), (2) balanced synonym sets per construct, (3) explicit inclusion of the baseline comparator construct, and (4) a rationale that states the intended precision–recall posture and documents exclusions.

Together, these syntheses show that retrieval posture and tool choice jointly shape what can be verified, replicated, and compared.

3.7. Reporting and Provenance Limitations

Across tools and runs, reporting biases emerged primarily as documentation and provenance gaps that constrain auditability independent of topical relevance. First, several runs produced incomplete or under-target lists or zero-yield outputs, creating selective visibility of evidence and forcing downstream analyses to rely on uneven top-*k* sets. Second, tool outputs varied in the explicitness and stability of rationales: some runs provided clear term-family justifications and exclusion intent, while others offered minimal explanation, limiting the interpretability of why particular records were retrieved and why others were absent. Third, browsing opacity (unclear indications of whether web browsing was active, which sources were accessed, and how ranking was computed) reduces provenance transparency and weakens claims about reproducibility and fair cross-tool comparison. Fourth, source substitution introduced a systematic bias toward non-peer-reviewed or non-indexed content presented in scholarly form (for example, organizational pages or web summaries replacing journal articles), which can inflate perceived coverage while reducing DOI-traceability under the study's standard. Collectively, these gaps operate as reporting biases because they differentially shape what is "observable" to the auditor and what can be verified, replicated, and attributed.

3.8. Confidence in the Audit Findings

Confidence in the audit findings is strongest for outcomes anchored to DOI-verified traceability and metadata integrity, because these outcomes were coded through direct resolution checks and bibliographic-field matching against authoritative records. Confidence is moderate for run-to-run drift estimates, because drift is sensitive to top-*k* truncation,

under-yield runs, and tool-specific ranking volatility; nevertheless, the observed turnover patterns are directionally consistent with the documented instability in identifier assignment and output composition. Confidence is currently provisional for bounded alignment with the librarian reference baseline, because the overlap analysis requires the finalized librarian target set and its DOI-verified identifiers; until that set is locked and overlap is computed, any alignment claims would be under-specified. Accordingly, the evidence base supports robust conclusions about integrity and reproducibility under a DOI-required standard, while claims about comparative relevance relative to librarian judgment remain pending the completion of the target-set overlap.

Taken together, the Results indicate that tool choice and retrieval posture materially affect traceability, metadata integrity, and run-to-run stability under a DOI-required standard.

4. Discussion

This audit yielded three core findings about retrieval posture and auditability. First, natural-language retrieval frequently produced non-verifiable or incorrectly matched citations, undermining traceability and metadata integrity under student-facing conditions. Second, Boolean translation with database execution increased the proportion of DOI-verifiable records and improved run completion relative to natural-language prompting. Third, run-to-run drift persisted across tools and postures, indicating that per-item citation correctness does not guarantee reproducible evidence sets.

Building on these results, the Discussion interprets tool choice and retrieval posture as workflow-level governance variables shaping traceability, metadata integrity, and reproducibility. The analysis extends prior work on evidence synthesis automation and AI-enabled review support by shifting evaluation away from speed or convenience toward DOI-level verifiability and bibliographic integrity in student-facing inquiry (Bolaños et al., 2024; Bond et al., 2024; Fütterer et al., 2026; Guo et al., 2024).

Recommendation for practice. When students need auditable and reproducible evidence sets, Boolean translation followed by database execution should be the default retrieval posture. In the present audit, this posture improved run completion and increased the proportion of DOI-verifiable records relative to natural-language prompting, making it a more defensible workflow for accountable inquiry. By contrast, natural-language retrieval is better treated as a provisional ideation aid for term discovery, concept expansion, and early orientation. Its outputs should not be used as final evidence lists unless each item is independently verified at the DOI and metadata level. Because drift persisted even under the more traceable posture, reproducibility also requires query logging, database identification, sorting-rule capture, and preserved export sets.

4.1. Accountability as a Workflow Property, Not a Tool Trait

The audit distinguishes perceived helpfulness from workflow accountability. Evidence synthesis guidance treats transparent search strategies and verifiable records as necessary conditions for defensible syntheses (Grant & Booth, 2009). In the audit, tools often produced fluent, plausible citations while degrading accountability through DOI wrong-match, non-resolving DOIs, and instability in core bibliographic fields. These failure modes align with broader concerns that AI systems can generate high-plausibility outputs that shift the burden from retrieval to verification, while obscuring provenance (Beer, 1984). Under a DOI-required verification standard (ISO, 2025), DOI wrong-match is the highest integrity risk because it simulates verification while redirecting the user to the wrong work. This pattern supports treating traceability (DOI resolvability) and metadata integrity (correct DOI–work match and field accuracy) as distinct constructs that require explicit evaluation, rather than assuming DOI presence implies correctness.

4.2. Retrieval Posture as a Governance Lever

Boolean translation and Scopus execution improved completion and increased the proportion of auditable records relative to natural-language prompting. This result is consistent with long-standing information retrieval practice, where explicit query logic, controlled vocabularies, and transparent constraints make search strategies auditable artifacts rather than implicit conversations. It also resonates with recent calls for transparency and rigor in educational AI adoption, which emphasize that governance should specify checkable process requirements rather than relying on tool assurances. Natural-language prompting approximates realistic student behavior but exposes students to opaque ranking, undocumented filtering, and higher integrity failure rates that are difficult to detect without systematic auditing. Boolean-mode retrieval constrains the search space through explicit constructs and field tags, thereby increasing auditability. However, Boolean quality varies at the construct level. When queries omit the librarian baseline comparator, they remain syntactically valid but become conceptually misaligned with the study question, increasing the likelihood of retrieving generic AI-in-education or automation papers rather than benchmark-comparative workflow evidence. Institutions should therefore teach Boolean translation as a method competency and require it as a documented compliance artifact, not treat it as an automatic safeguard.

4.3. Reproducibility Requires Stability

Run-to-run drift indicates that high per-item DOI correctness does not guarantee reproducibility of the retrieved evidence set (not only the accuracy of individual citations). Even when a tool returns mostly DOI-correct citations, repeated executions can return materially different top-*k* sets. This finding complicates the common assumption that "verified citations" imply "repeatable searches." In evidence synthesis traditions, reproducibility requires that the search strategy, execution conditions, and selection steps be sufficiently documented so that others can re-run and approximate the evidence base. The present results indicate that reproducibility depends on both the explicitness of the search protocol and the stability of ranking and selection heuristics. From a governance standpoint, drift requires students to archive the exact query, execution date, database, sorting rule, and exported results, rather than treating the evidence set as a stable product of the topic alone.

4.4. Metadata Integrity as an Equity and Attribution Issue

Bibliographic field instability has equity implications, particularly for authors whose names include multiple given names and or multiple family names. Misparsing names distorts attribution, impairs deduplication, and propagates errors through citation managers and screening workflows. This concern aligns with data governance and FAIR-oriented arguments that metadata quality is a precondition for discoverability, reuse, and responsible synthesis (Guillen-Aguinaga et al., 2025). The issue is not cosmetic. It shapes how scholarship is counted, discovered, and credited, and it can bias student synthesis by fragmenting author identities and obscuring collaboration networks. Institutions should therefore treat correct author parsing and APA 7 compliance as integrity requirements, and they should teach students to verify author strings against authoritative records, particularly when using AI-generated citations.

4.5. Implications for Student Epistemic Agency and Assessment Validity

The audit reframes AI-assisted literature review as a test of epistemic agency. Students do not only retrieve sources. They justify why a source belongs, verify that it exists as cited, and document how it was found. Recent syntheses of AI in higher education emphasize that rigor and ethics must be operationalized in practice, particularly when tools mediate core

academic work. In the present audit, tool-generated citations that appear complete but fail verification can shift student labor away from reasoning and toward unproductive repair, or toward uncritical acceptance. This dynamic risks undermining assessment validity because assignments can inadvertently measure tool behavior rather than student understanding of search logic, review design, and verification. A governance-ready response is to assess process artifacts, not only the final reference list. Courses can require a minimal audit trail that includes the query, database, filters, exported set, DOI verification notes, and a brief inclusion rationale tied to research-question constructs.

4.6. Institutional Governance and Librarian Integration

The findings support a governance model where librarians define baseline expectations for search strategy quality and instructors and research offices enforce documentation standards for AI-assisted workflows. This division of labor aligns with evidence synthesis norms that treat search strategy construction and reporting as specialized competencies, often strengthened by librarian involvement. Three institutional actions follow directly from the Results. First, standardize a student-facing verification checklist. Any AI-generated citation used in coursework should pass a verification protocol, including DOI resolution and metadata match, or a documented alternative identifier check when no DOI exists. Second, require protocol disclosure for AI use. Students should disclose the tool, prompt, browsing status, database execution conditions, and post-processing. Third, embed librarian benchmark alignment. A librarian benchmark can support instructionally meaningful feedback about systematic omissions and false positives.

4.7. Reconciliation with Analytic Expectations

The study proceeded with three audit expectations derived from the research question and evidence synthesis practice. First, the team expected Boolean-mode retrieval executed in Scopus to increase completion and DOI-traceability relative to conversational retrieval. This expectation was supported. Second, the team expected Boolean constraints to reduce run-to-run drift by stabilizing construct specification. This expectation was partially supported, with moderate improvements for some tools but persistent instability for others, indicating that query constraints do not fully control ranking sensitivity. Third, the team expected tool families to differ in bibliographic integrity under identical prompts. This expectation was supported, with meaningful between-tool variation in DOI mismatch, DOI resolvability, and APA-field stability. These reconciliations clarify that the primary governance signal is not whether a tool can produce correct citations in a single run, but whether a workflow can reliably generate verifiable and repeatable evidence sets under teachable rules.

4.8. Limitations and Next Steps

Two primary limitations bound interpretation. First, the audit evaluates tool outputs under a fixed canonical prompt and a top-k capture rule. Alternative prompts, model settings, database filters, or ranking choices may materially alter yield, DOI traceability, metadata integrity, and run-to-run drift. Second, overlap and alignment estimates depend on the librarian benchmark, which is contingent on expert judgment and database access conditions. While librarian augmentation strengthens the methodological baseline by anchoring the study in established information-retrieval and evidence-synthesis practice, the benchmark remains contingent on expert judgment, search system access (for example, availability of Scopus and Web of Science), and the librarian's domain framing. These boundary conditions limit generalizability and should be treated explicitly when interpreting overlap, precision, and recall.

Next steps include expanding overlap analyses at the DOI-level identity across all AI runs, reporting bounded precision and bounded recall relative to the librarian target set under the same top-*k* constraint, and preregistering prompt variants to test how prompt controls affect drift and integrity outcomes. A further extension is to replicate the Boolean execution stage across additional databases and interfaces (for example, Web of Science and ERIC) to test whether stability and traceability improve under alternative ranking, indexing, and filtering regimes. Finally, scaling the audit will require standardizing benchmark construction (for example, multi-librarian adjudication or consensus procedures) and documenting access dependencies to support reproducibility across institutions.

4.9. Practical Recommendations

For student-facing workflows, the audit supports four minimal requirements: explicit construct definition, database-executed structured search, DOI and metadata verification, and archival preservation of the full retrieval trail. These requirements treat AI as a governed workflow component rather than as a substitute for methodological understanding.

5. Conclusions

In conclusion, this study shows that AI-assisted literature search workflows are best evaluated as accountable inquiry rather than as generic tool performance. The audit protocol offers instructors, librarians, and research administrators a practical mechanism to govern AI use in student-facing work by foregrounding reproducibility and transparency at the level of citations and evidence trails. The protocol reframes evaluation through academic rigor and operationalizes DOI-based traceability, metadata integrity, and run-to-run reproducibility across tools and retrieval postures. The principal contribution is methodological and practical. Methodologically, the work specifies auditable constructs, observable failure codes, and a verification escalation procedure that can be replicated and extended. Practically, it reframes AI-mediated literature searching as a workflow that must preserve an evidence trail suitable for classroom assessment and institutional oversight.

Tool choice and retrieval posture materially shape whether students can demonstrate verifiable sourcing. Natural-language outputs frequently produced integrity failures that are difficult to detect without structured auditing, including DOI wrong-match errors that mimic verification while redirecting users to the wrong record. Boolean translation and database execution increased run completion and shifted retrieval toward DOI-bearing records, improving auditability. However, higher per-item DOI correctness did not guarantee stability of the retrieved evidence set, indicating that reproducibility requires explicit protocol capture beyond citation accuracy. These implications extend to educational practice: instructors can improve assessment validity by requiring process artifacts, including the executed query, database, and sorting rule, exported result sets, and DOI verification notes. Institutions can strengthen governance by using librarian baselines as reference benchmarks for search strategy quality and transparent reporting.

Collectively, these findings support institutional policies that treat AI-assisted literature searching as a documented method choice with verification obligations, anchored to librarian-defined standards for search quality and transparent reporting.

Decision rule. In student-facing inquiry, natural-language prompting should be limited to exploratory term generation and early topic orientation, whereas the evidence set used for screening, coding, and citation should be derived from database-executed Boolean retrieval with documented query logic and DOI-level verification.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/aieduc2020008/s1. Supplementary files follow the journal's "S" naming convention (for example, Figure S1, Table S1, Dataset S1, and Document S1). File names

and labels below have been aligned with the submitted supplementary archive. Item counts may be updated after eligibility screening and final inclusion decisions are locked. An archival copy of the supplementary materials is also available in OSF. Box S1. Canonical prompt (verbatim): Locked research question, task constraints, allowed deviations log, and drift operationalization rules. Dataset S1. Zotero expert resources list (PDF): LIS expert baseline resource list corresponding to Dataset S7 (n = 27), including bibliographic metadata and identifiers. Dataset S2. Zotero export of retrieved records (RIS): Export of all captured records and metadata fields prior to staged screening. Dataset S3. Identification All Runs (CSV) (n = 212): Zotero group export after initial deduplication, including provenance fields such as tool, run ID, and retrieval posture. Dataset S4. Screening Cleaned (CSV) (n = 170): Post-CSV re-deduplicated unique-record dataset, including provenance fields and deduplication flags. Dataset S5. Full-text eligibility (CSV) (n = 100): Records advanced to full-text assessment, including inclusion and exclusion codes and adjudication status. Dataset S6. Included Articles (CSV) (n = 37): Final included set after quality appraisal, with quality appraisal outcomes and analysis-ready fields. Dataset S7. Expert Sources (CSV) (n = 27): LIS expert contribution set, including "Extra" source tags (Google Scholar, Scopus, Web of Science) and DOI-verification status. Dataset S8. Zotero Included Articles (PDF): Final included selection with bibliographic metadata and identifiers. Document S1. Full run outputs (PDF): All runs, including tool responses, extracted citations, and Scopus-format Boolean queries. Document S2. DSSC (PDF): Disclosure of Support Statement and Statement of Contributions. Document S3. DMSP (PDF): Data Management and Sharing Plan. Document S4. Traceability and Screening Codebook (PDF): Operational definitions, screening rules, and traceability failure classifications used in the audit. Document S5. PRISMA 2020 abstract Checklist. Document S6. PRISMA 2020 Checklist. Figures S1 and S2. Baseline, expert-seeded diagnostic VOSviewer overlays for Dataset S7 (n = 27): Figure S1 shows the keyword co-occurrence overlay; Figure S2 shows the co-authorship overlay. Table S1. Run log and captured outputs by tool and retrieval posture: Natural-language and Boolean runs, including run IDs, completion status, and item counts for Gemini (institutional), ChatGPT free tier, Perplexity (Academic configuration), and ChatGPT paid tier (Extended Thinking). Table S2. Traceability failure modes: Operational definitions and classification rules for traceability failure modes observed during auditing.

Author Contributions: Conceptualization, C.L. and M.K.; methodology, C.L.; software, C.L.; validation, C.L. and M.K.; formal analysis, C.L.; investigation, C.L. and M.K.; resources, C.L.; data curation, C.L.; writing—original draft preparation, C.L.; writing—review and editing, C.L. and M.K.; visualization, C.L.; supervision, C.L.; project administration, C.L.; funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study involved document-based analysis of publicly available bibliographic records and DOI resolution outcomes and did not include human participants, identifiable private information, or interventions; therefore, Institutional Review Board review was not sought.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated and analyzed in this study (canonical prompt, run logs and captured outputs, Scopus Boolean strings, Zotero exports, deduplicated CSV, traceability codebook (Supplementary Material: Document S4), and calculation sheets/scripts used to reproduce tables) will be made publicly available as Supplementary Materials at the OSF https://osf.io/htdya/overview?view_only=6dec76cf26174abbb77361b5a11bfc4a, accessed on 15 March 2026. A Data Management and Sharing Plan (DMSP) will document versions, access points, and downstream project outputs; https://dmphub.uc3prd.cdlib.net/dmps/10.48321/D14AEBB6CF, accessed on 15 March 2026.

Acknowledgments: The authors acknowledge the administrative and technical support provided through institutional facilities and library access that enabled database searching, record management, and documentation of the audit workflow. All support, tool use, and contributions, including any

GenAI assistance used during study design, data collection, analysis, interpretation, or manuscript preparation, are fully disclosed in the Disclosure of Support and Statement of Contributions (DSSC) (Document S2) included in the Supplementary Materials. During the preparation of this study, the authors used generative AI tools (as specified in the DSSC) to generate candidate citations, propose Scopus-style Boolean queries, and draft explanatory rationales as part of the audited workflows. The authors reviewed and edited all tool outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

PRISMA: Although the study follows PRISMA 2020 reporting logic to document corpus identification, screening, and inclusion, it should be understood as a review-informed empirical audit of AI-assisted retrieval workflows rather than as a standalone systematic literature review. OSF registration: Open Science Framework (OSF), Registration DOI: 10.17605/OSF.IO/U8NHT; Registration record: https://osf.io/u8nht/overview, accessed on 15 March 2026.

Amendments to Protocol: No amendments reported at time of registration. Any future changes to the protocol or registration record will be documented in OSF versioning and described in the manuscript if applicable.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
APA 7	American Psychological Association, 7th edition citation style
CGT	Constructivist Grounded Theory
CSV	Comma-separated values (tabular data file format)
DMSP	Data Management and Sharing Plan
DOI	Digital Object Identifier
DSSC	Disclosure of Support Statement of Contributions
ERIC	Education Resources Information Center (bibliographic database)
GenAI	Generative artificial intelligence
GPT	Generative Pre-trained Transformer
IR	Information retrieval
LIS	Library and Information Science
LLM	Large language model
NJIT	New Jersey Institute of Technology
NO_DOI	Traceability code: no DOI provided
OSF	Open Science Framework
PDF	Portable Document Format
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RIS	Research Information Systems format (bibliographic interchange file format)
SLR	Systematic literature review
TITLE-ABS-KEY	Scopus field tag: Title, Abstract, Keywords
VOSviewer, version 1.8.0_202	Bibliometric mapping and network visualization software

Appendix A

Appendix A.1. Boolean Strings: Librarian

Source	Full String	String Used
Librarian 3 Runs	TITLE-ABS-KEY ((“AI” OR “artificial intelligence” OR “LLM” OR “large language model”) AND (“literature review” OR “systematic review” OR “literature survey”))	TITLE-ABS-KEY ((“AI” OR “artificial intelligence” OR “LLM” OR “large language model”) AND (“literature review” OR “systematic review” OR “literature survey”))

Appendix A.2. Boolean Strings: Gemini

Source	Full String	String Used
Gemini (institutional). Run 1	TITLE-ABS-KEY ((“artificial intelligence” OR “AI-assisted” OR “LLM” OR “ChatGPT”) AND (“literature review” OR “systematic review” OR “evidence synthesis”) AND (“traceability” OR “metadata integrity” OR “reproducibility” OR “rigor”) AND (“undergraduate” OR “graduate” OR “student”) AND (“librarian” OR “information literacy” OR “manual search”)))	TITLE-ABS-KEY ((“artificial intelligence” OR “AI-assisted” OR “LLM” OR “ChatGPT”) AND (“literature review” OR “systematic review” OR “evidence synthesis”) AND (“traceability” OR “metadata integrity” OR “reproducibility” OR “rigor”) AND (“undergraduate” OR “graduate” OR “student”) AND (“librarian” OR “information literacy” OR “manual search”)))
Gemini (institutional). Run 2	TITLE-ABS-KEY ((“artificial intelligence” OR “AI” OR “LLM” OR “ChatGPT”) AND (“literature review” OR “systematic review” OR “evidence synthesis”) AND (“traceability” OR “metadata” OR “reproducibility” OR “provenance”) AND (“undergraduate” OR “graduate” OR “student”)))	TITLE-ABS-KEY ((“artificial intelligence” OR “AI” OR “LLM” OR “ChatGPT”) AND (“literature review” OR “systematic review” OR “evidence synthesis”) AND (“traceability” OR “metadata” OR “reproducibility” OR “provenance”) AND (“undergraduate” OR “graduate” OR “student”)))

Appendix A.3. Boolean Strings: ChatGPT Free Tier

Source	Full String	String Used
ChatGPT free tier. Run 3	(TITLE-ABS-KEY(“artificial intelligence” OR “AI” OR “machine learning” OR “large language model”) AND TITLE-ABS-KEY(“systematic review” OR “literature review” OR “evidence synthesis”) AND TITLE-ABS-KEY(“traceability” OR “metadata” OR “reproducibility” OR “integrity” OR “workflow”))	(TITLE-ABS-KEY(“artificial intelligence” OR “AI” OR “machine learning” OR “large language model”) AND TITLE-ABS-KEY(“systematic review” OR “literature review” OR “evidence synthesis”) AND TITLE-ABS-KEY(“traceability” OR “metadata” OR “reproducibility” OR “integrity” OR “workflow”))
ChatGPT free tier. Run 4	(TITLE-ABS-KEY((“artificial intelligence” OR AI OR “machine learning” OR “large language model*”) AND (“literature review” OR “systematic review” OR “evidence synth*”) AND (traceability OR reproducib* OR “metadata integrity” OR transparency OR “search strategy” OR librarian* OR workflow)) AND (LIMIT-TO(DOCTYPE, “ar”) OR LIMIT-TO(DOCTYPE, “re”))	(TITLE-ABS-KEY((“artificial intelligence” OR AI OR “machine learning” OR “large language model*”) AND (“literature review” OR “systematic review” OR “evidence synth*”) AND (traceability OR reproducib* OR “metadata integrity” OR transparency OR “search strategy” OR librarian* OR workflow))

Appendix A.4. Boolean Strings: Perplexity

Source	Full String	String Used
Perplexity (Academic configuration). Run 5	TITLE-ABS-KEY((“artificial intelligence” OR “machine learning” OR “large language model*” OR “LLM” OR “AI-assisted” OR “AI-assisted tool*” OR “automation tool*” OR “systematic review automation” OR “semi-automated” OR “computer-assisted”) AND (“literature review*” OR “systematic review*” OR “evidence synthesis” OR “scoping review*” OR “rapid review*”) AND (traceab* OR “provenance” OR “audit trail” OR “workflow log*” OR “process documentation” OR reproducib* OR replicab* OR “metadata” OR “metadata quality” OR “data lineage” OR “search strategy transparenc*” OR “PRISMA-trAIce”)) AND TITLE-ABS-KEY((librarian* OR “information specialist*” OR “information professional*” OR “mediated search*” OR “expert search*”) OR (student* OR undergraduate* OR postgraduate* OR “graduate student*” OR “novice reviewer*” OR “novice researcher*”)) AND (LIMIT-TO(DOCTYPE, “ar”) OR LIMIT-TO(DOCTYPE, “re”)) AND PUBYEAR > 2015	TITLE-ABS-KEY((“artificial intelligence” OR “machine learning” OR “large language model*” OR “LLM” OR “AI-assisted” OR “AI-assisted tool*” OR “automation tool*” OR “systematic review automation” OR “semi-automated” OR “computer-assisted”) AND (“literature review*” OR “systematic review*” OR “evidence synthesis” OR “scoping review*” OR “rapid review*”) AND (traceab* OR “provenance” OR “audit trail” OR “workflow log*” OR “process documentation” OR reproducib* OR replicab* OR “metadata” OR “metadata quality” OR “data lineage” OR “search strategy transparenc*” OR “PRISMA-trAIce”)) AND TITLE-ABS-KEY((librarian* OR “information specialist*” OR “information professional*” OR “mediated search*” OR “expert search*”))

Source	Full String	String Used
Perplexity (Academic configuration). Run 6	(TITLE-ABS-KEY(“systematic review*” OR “literature review*” OR “evidence synthesis” OR “scoping review*”)) (TITLE-ABS-KEY(“artificial intelligence” OR “AI-assisted” OR “AI based” OR “machine learning” OR “large language model*” OR “LLM*” OR “automation tool*” OR “screening tool*” OR “Rayyan” OR “DistillerSR” OR “RobotAnalyst”)) AND (TITLE-ABS-KEY(reproducib* OR “metadata integrity” OR “metadata quality” OR traceab* OR “audit trail” OR “research transparency” OR “transparent reporting”)) AND (TITLE-ABS-KEY(librarian* OR “information specialist*” OR “library professional*” OR “library science” OR “academic library”)) AND (LIMIT-TO(DOCTYPE, “ar”) OR LIMIT-TO(DOCTYPE, “re”))	(TITLE-ABS-KEY(“systematic review*” OR “literature review*” OR “evidence synthesis” OR “scoping review*”)) (TITLE-ABS-KEY(“artificial intelligence” OR “AI-assisted” OR “AI based” OR “machine learning” OR “large language model*” OR “LLM*” OR “automation tool*” OR “screening tool*” OR “Rayyan” OR “DistillerSR” OR “RobotAnalyst”)) AND (TITLE-ABS-KEY(reproducib* OR “metadata integrity” OR “metadata quality” OR traceab* OR “audit trail” OR “research transparency” OR “transparent reporting”)) AND (TITLE-ABS-KEY(librarian* OR “information specialist*” OR “library professional*” OR “library science” OR “academic library”))

Appendix A.5. Boolean Strings: ChatGPT Paid Tier (Extended Thinking)

Source	Full String	String Used
ChatGPT paid tier (Extended Thinking). Run 7	TITLE-ABS-KEY((“systematic review” OR “scoping review” OR “evidence synthesis” OR “literature review”) AND (“artificial intelligence” OR “machine learning” OR “large language model*” OR automation OR “AI-assisted” OR “human-in-the-loop” OR textmining OR “screening tool*”) AND (traceab* OR reproducib* OR “audit trail” OR provenance OR metadata OR “search strateg*” OR deduplicat* OR “reference management”) AND (librarian* OR “information specialist*” OR student* OR undergraduat* OR graduat* OR thesis OR dissertation))	TITLE-ABS-KEY((“systematic review” OR “scoping review” OR “evidence synthesis” OR “literature review”) AND (“artificial intelligence” OR “machine learning” OR “large language model*” OR automation OR “AI-assisted” OR “human-in-the-loop” OR textmining OR “screening tool*”) AND (traceab* OR reproducib* OR “audit trail” OR provenance OR metadata OR “search strateg*” OR deduplicat* OR “reference management”) AND (librarian* OR “information specialist*” OR student* OR undergraduat* OR graduat* OR thesis OR dissertation))
ChatGPT paid tier (Extended Thinking). Run 8	TITLE-ABS-KEY((“systematic review” OR “literature review” OR “evidence synthesis” OR “scoping review”) AND (“artificial intelligence” OR “machine learning” OR “large language model*” OR ChatGPT OR “active learning” OR “text mining” OR automation OR “semi-automated” OR “technology-assisted”) AND (traceab* OR reproducib* OR transparen* OR “audit trail” OR provenance OR “workflow” OR “search strateg*” OR metadata) AND (librarian* OR “information specialist*” OR “expert searcher*” OR “reference service*”))	TITLE-ABS-KEY((“systematic review” OR “literature review” OR “evidence synthesis” OR “scoping review”) AND (“artificial intelligence” OR “machine learning” OR “large language model*” OR ChatGPT OR “active learning” OR “text mining” OR automation OR “semi-automated” OR “technology-assisted”) AND (traceab* OR reproducib* OR transparen* OR “audit trail” OR provenance OR “workflow” OR “search strateg*” OR metadata) AND (librarian* OR “information specialist*” OR “expert searcher*” OR “reference service*”))

Appendix A.6. One-Page Student Checklist: Per Run

Audit of AI-Assisted Literature Search Outputs (Y/N)

Instructions. Use Y/N for each item. Complete A + B once per run. (Per-citation checks appear in Appendix A.2.)

Run ID: _____ Tool/Mode (free/paid/institutional): _____

Retrieval settings (model/version, mode, browsing on/off): _____

Date/Time: _____

A. Run Registration (complete once)

Criterion	Yes	No
1. Tool name and access mode recorded (free/paid, etc.).	☐	☐
2. Date and time recorded.	☐	☐
3. Full prompt saved verbatim.	☐	☐
4. Raw output saved verbatim.	☐	☐
5. Retrieval settings recorded (model, version, browsing on/off)	☐	☐

B. Reproducibility and Drift (complete once)

Criterion	Yes	No
6. Another student could rerun from your log (prompt + settings).	☐	☐
7. Second run completed and saved (same prompt).	☐	☐
8. Overlap check acceptable ($\geq$50% same items across run 1 vs. run 2).	☐	☐
9. If overlap is low (i.e., <50%), likely cause(s) are documented.	☐	☐
10. Export created (Zotero/EndNote/CSV) for audit trail.	☐	☐

Scoring (fill in)

Run compliance: A(1–5) ___/5 + B(6–10) ___/5

Records audited (Ra): ___ Record passes (Rp): ___ Pass rate Rp/Ra $\times$ 100 = ___%

Appendix A.7. One-Page Student Checklist: Per Citation

Instructions. Use Y/N for each item. Complete C–E per citation (or sample first 20 if instructed).

C. Traceability (per citation)

Criterion	Yes	No
11. DOI or stable ID explicitly provided (e.g., PMID).	☐	☐
12. If DOI is provided, it resolves (no error).	☐	☐
13. Source page matches title, first author, and additional authors.	☐	☐
14. If no DOI: marked **NO DOI** (no guessing).	☐	☐
15. Bibliographic page/abstract accessible for verification.	☐	☐

D. Metadata Integrity (per citation; verify against source page)

Criterion	Yes	No
16. Author list/order matches.	☐	☐
17. Publication year matches.	☐	☐
18. Venue (journal/proceedings) matches.	☐	☐
19. Title matches (ignoring capitalization).	☐	☐
20. Item type correct (peer-reviewed journal/proceedings).	☐	☐

E. Eligibility and Boundary (per citation)

Criterion	Yes	No
21. Peer-reviewed (not gray literature).	☐	☐
22. AI-assisted literature search/review (not only general AI).	☐	☐
23. Addresses ≥1 target outcome: traceability OR metadata integrity OR reproducibility.	☐	☐
24. If out-of-scope, excluded (not retained "just in case").	☐	☐

Record pass rule (recommended).
Exclude if 24 = Y.
Otherwise, Pass if 11, 15, 16, 17, 18, 19, 20, 21, 22, 23 = Y and [(12 and 13 = Y) or (14 = Y)].
Confidence tier (recommended). High more than 11 (≥80% pass + run compliant)/Moderate 7 to 10 (50–79%)/Low 6 or less (<50% or noncompliant).
Artifacts to submit. Prompt + raw outputs (Run 1 & Run 2), verification notes, export (Zotero/CSV), completed checklists.

Appendix A.8. One-Page Student Masterlist

Master Validity and Quality Screen (Y/N)
Have you validated the source?
☐ Yes ☐ No
(Yes, only if author, year, title, venue, and DOI/NO DOI + stable URL are verified against the source page.)
Is it available (can you access it and verify the DOI)?
☐ Yes ☐ No
(Yes, only if you can open the DOI landing page or stable record page, and access at least the abstract/bibliographic record; ideally, full text is downloadable via OA/library/ILL.)
Can you summarize the content in the context of your intended use in your work?
☐ Yes ☐ No
(Yes, only if you can write 2–3 sentences: what it argues/finds + how you will use it—background, theory, methods, evidence, or benchmark.)
Can you justify the source's authority/quality for your purpose?
☐ Yes ☐ No
(Yes, only if it is credible for your review posture: peer-reviewed venue or explicitly allowed gray literature; venue is reputable; methods/claims are appropriate; timeliness/foundational value fits your time window.)
Decision rule:
Include if all four = Yes.
Exclude if 1 or 2 = No.
Hold/Review if 1–2 = Yes, but 3 or 4 = No (needs reading, access, or quality justification).

Notes

[1] Staged screening and appraisal were aligned with the information level of the 4i Framework for Advancing Communication and Trust (4i FACT), emphasizing amplification of factual information, mitigation of misinformation, and credibility assessment across sources.

[2] Intermediate Zotero library after first-pass deduplication, see Dataset S3 (n = 212); prior to CSV re-deduplication and LIS augmentation.

[3] Although ANOCVA offers a principled approach for comparing clustering structures (Patriota et al., 2018), the current corpus and study scope do not provide sufficient independent clustering instances to support a full ANOCVA application with stable inference. Accordingly, this study specifies the ANOCVA comparison as an intended extension and treats the present audit as a baseline protocol for scalable replication. After completion of the planned pilot implementation with approximately 50 students,

the protocol will be expanded to a larger cohort (target n ≈ 500) to enable robust comparative analyses of clustering structures and workflow effects under student-use conditions.

[4] Note that 0 ≤ overlap (A, B) ≤ 1. If set A is a subset of B or the converse, then the overlap coefficient is equal to 1.

[5] The Jaccard index (Jaccard, 1901), originally termed the coefficient de communauté (coefficient of community), is also known as the ratio of verification (Murphy, 1996) and, in meteorology, the critical success index. An independent formulation was developed by Tanimoto (1958).

References

Abdullah, K. K. A. (2016). Effect of similarity measures of terms in information retrieval system. *Advances in Multidisciplinary & Scientific Research*, *2*(4), 219–228.

Baker, G., & Caton, L. (Eds.). (2025). *AI-powered pedagogy and curriculum design: Practical insights for educators*. Routledge. [CrossRef]

Ballesteros-Ballesteros, V. A., & Zárate-Torres, R. A. (2025). Mapping the conceptual structure of research on open innovation in university–industry collaborations: A bibliometric analysis. *Frontiers in Research Metrics and Analytics*, *10*, 1693969. [CrossRef] [PubMed]

Bandara, W., Furtmueller, E., Gorbacheva, E., Miskon, S., & Beekhuyzen, J. (2015). Achieving rigor in literature reviews: Insights from qualitative data analysis and tool-support. *Communications of the Association for Information Systems*, *37*(1), 154–204. [CrossRef]

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & van der Linden, S. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, *8*(1). [CrossRef]

Beer, S. (1984). The viable system model: Its provenance, development, methodology and pathology. *Journal of the Operational Research Society*, *35*(1), 7–25. [CrossRef]

Bolaños, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: Opportunities and challenges. *Artificial Intelligence Review*, *57*(10), 259. [CrossRef]

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(1), 4. [CrossRef]

Bukar, U. A., Sayeed, M. S., Razak, S. F. A., Yogarayan, S., Amodu, O. A., & Mahmood, R. A. R. (2023). A method for analyzing text using VOSviewer. *MethodsX*, *11*, 102339. [CrossRef]

Buyya, R., Calheiros, R. N., & Vahid Dastjerdi, A. (2016). *Big data* (1st ed.). Elsevier. [CrossRef]

Charmaz, K. (2017). The power of constructivist grounded theory for critical inquiry. *Qualitative Inquiry*, *23*(1), 34–45. [CrossRef]

Chaudhry, M. A., Cukurova, M., & Luckin, R. (2022). A transparency index framework for AI in education. In *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners' and doctoral consortium*. Springer.

Clark, J., McFarlane, C., Cleo, G., Ramos, C. I., & Marshall, S. (2021). The impact of systematic review automation tools on methodological quality and time taken to complete systematic review tasks: Case study. *JMIR Medical Education*, *7*(2), e24418. [CrossRef]

Deroncele-Acosta, A. (2025). QR method: A step-by-step guide to writing a narrative review. *Islas*, *67*(212), e1610.

Eusebio, E. J. G., Baldera, P., Patiam, A. M., Villanueva, E. R., Gaa, N. A., Solis, A. M. F., Soriano, M. L. C., & Ribon, A. L. (2025). AI in the classroom: A systematic review of barriers to educator acceptance. *International Journal of Learning, Teaching and Educational Research*, *24*(9), 126–147. [CrossRef]

Fink, A. G. (2019). *Conducting research literature reviews: From the internet to paper* (5th ed.). SAGE Publications, Inc.

Franco-Pereira, A. M., Nakas, C. T., Reiser, B., & Carmen Pardo, M. (2021). Inference on the overlap coefficient: The binormal approach and alternatives. *Statistical Methods in Medical Research*, *30*(12), 2672–2684. [CrossRef]

Frey, B. B. (2018). *The sage encyclopedia of educational research, measurement, and evaluation* (1st ed.). SAGE Publications, Inc. [CrossRef]

Fütterer, T., Campos, D. G., Gfrörer, T., Lavelle-Hill, R., Murayama, K., & Scherer, R. (2026). AI tools for systematic literature reviews and meta-analyses in educational psychology: An overview and a practical guide. *Learning and Individual Differences*, *126*, 102849. [CrossRef]

Gough, D., Oliver, S., & Thomas, J. (2021). *An introduction to systematic reviews* (2nd ed.). SAGE Publications, Inc.

Göndöcs, D., Horváth, S., & Dörfler, V. (2025). Uncovering the dynamics of human-AI hybrid performance: A qualitative meta-analysis of empirical studies. *International Journal of Human-Computer Studies*, *205*, 103622. [CrossRef]

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91–108. [CrossRef]

Guillen-Aguinaga, M., Aguinaga-Ontoso, E., Guillen-Aguinaga, L., Guillen-Grima, F., & Aguinaga-Ontoso, I. (2025). Data quality in the age of AI: A review of governance, ethics, and the fair principles. *Data*, *10*(12), 201. [CrossRef]

Guo, E., Gupta, M., Deng, J., Park, Y.-J., Paget, M., & Naugler, C. (2024). Automated paper screening for clinical reviews using large language models: Data analysis study. *Journal of Medical Internet Research*, *26*(1), e48996. [CrossRef] [PubMed]

Hardie, P., Darley, A., Derwin, R., Eustace-Cook, J., Kearns, S., Brien, B. M., Siddiquee, A., Zheng, D., & Mooney, M. (2025). Applications, attitudes and ethical considerations of generative artificial intelligence (Gen AI) in nursing education: A scoping review. *BMC Nursing*, *25*(1), 148. [CrossRef] [PubMed]

Haukås, Å., Bjørke, C., & Dypedahl, M. (2018). *Metacognition in language learning and teaching* (1st ed.). Routledge Studies in Applied Linguistics. Taylor & Francis.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. d. P. (2014). *Metodología de la investigación, sexta edición, impreso* (6th ed.). McGraw-Hill/Interamericana Editores, S.A. de C.V.

Holst, D., Moenck, K., Koch, J., Schmedemann, O., & Schüppstuhl, T. (2025). Transparent reporting of AI in systematic literature reviews: Development of the PRISMA-trAIce checklist. *Jmir Ai*, *4*, e80247. [CrossRef] [PubMed]

Höpner, A., & Uckelmann, D. (2024). Implementing generative AI in academic writing seminars. In *EDULEARN24 proceedings* (pp. 4998–5005). International Academy of Technology, Education and Development (IATED).

ISO. (2025). *Information and documentation—Digital object identifier system* (ISO 26324:2012(en)). International Organization for Standardization.

ISO 15836. (2017). *Information and documentation—The Dublin core metadata element set—Part 1: Core elements* (ISO 15836-1:2017). International Organization for Standardization.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de La Société Vaudoise Des Sciences Naturelles*, *37*(142), 547. [CrossRef]

Jesson, J. (2011). *Doing your literature review, first edition: Traditional and systematic techniques* (1st ed.). SAGE Publications Ltd.

Kajan, K., Shi, W., Wanatowski, D., Kajan, K., Shi, W., & Wanatowski, D. (2025). STEM undergraduates' perceptions of AI chatbots: A cross-sectional descriptive survey. *AI in Education*, *1*(1), 4. [CrossRef]

Kallmes, K. M., Thurnham, J., Sauca, M., Tarchand, R., Kallmes, K. R., & Holub, K. J. (2025). Human-in-the-loop artificial intelligence system for systematic literature review: Methods and validations for the AutoLit review software. *Cochrane Evidence Synthesis and Methods*, *3*(6), e70059. [CrossRef]

Kannan, S., Karuppusamy, S., Nedunchezhian, A., Venkateshan, P., Wang, P., Bojja, N., & Kejariwal, A. (2016). Big data analytics for social media. In *Big data* (pp. 63–94). Elsevier.

Kotu, V., & Deshpande, B. (2019). *Data science: Concepts and practice*. Morgan Kaufmann.

Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. *Patterns*, *2*(9), 100322. [CrossRef]

Li, J., Yan, Y., & Zeng, X. (2025). Exploring artificial intelligence in inclusive education: A systematic review of empirical studies. *Applied Sciences*, *15*(23), 12624. [CrossRef]

Lochmiller, C., & Lester, J. N. (2016). *An introduction to educational research: Connecting methods to practice* (1st ed.). SAGE Publications, Inc.

Mabirizi, V., Katushabe, C., Muhoza, G., & Rugasira, J. (2025). A systematic review of the impact of generative AI on postgraduate research: Opportunities, challenges, and ethical implications. *Discover Artificial Intelligence*, *5*(1), 238. [CrossRef]

Montoya, F. G., Alcayde, A., Baños, R., & Manzano-Agugliaro, F. (2018). A fast method for identifying worldwide scientific collaborations using the scopus database. *Telematics and Informatics*, *35*(1), 168–185. [CrossRef]

Murphy, A. H. (1996). *The finley affair: A signal event in the history of forecast verification*. Available online: https://journals.ametsoc.org/view/journals/wefo/11/1/1520-0434_1996_011_0003_tfaase_2_0_co_2.xml (accessed on 24 January 2026).

Ndéla Marone, R. M., & Mbengue, M. (2025). Chatbots and artificial intelligence to support digital university libraries in Africa: Opportunities and challenges. In L. Yang, & A. Salaz (Eds.), *Digital libraries across continents* (pp. 72–92). Taylor and Francis. [CrossRef]

Omeh, C. B., & Ayanwale, M. A. (2025). Artificial intelligence meets PBL: Transforming computer-robotics programming motivation and engagement. *Frontiers in Education*, *10*, 1674320. [CrossRef]

Pai H., A., Pathak, G. R., Agarwal, J., S, C., Parinam, S., & Gupta, S. (2025). Integrating artificial intelligence in stem classrooms: A mixed-methods study of models, outcomes, and ethical challenges. In *2025 3rd international conference on data science and network security (ICDSNS)* (pp. 1–5). IEEE.

Park, S. G. (2025). AI and systematic reviews: Can AI tools replace librarians in the systematic search process? *Science and Technology Libraries*, 1–22. [CrossRef]

Park, S. G., Carroll, M., Esteve, L. M. A., & Singh, K. (2024). Exploring generative AI and natural language processing to develop search strategies for systematic reviews. In *American Society for Engineering Education annual conference and exposition, conference proceedings*. American Society for Engineering Education.

Patriota, A. G., Calebe Vidal, M., Caetano de Jesus, D. A., & Fujita, A. (2018). ANOCVA: A nonparametric statistical test to compare clustering structures. In *Theoretical and applied aspects of systems biology* (pp. 113–125). Springer.

Pawliuk, C., Cheng, S., Zheng, A., & Siden, H. (2024). Librarian involvement in systematic reviews was associated with higher quality of reported search methods: A cross-sectional survey. *Journal of Clinical Epidemiology*, *166*, 111237. [CrossRef] [PubMed]

Post, C., Sarala, R., Gatrell, C., & Prescott, J. E. (2020). Advancing theory with review articles. *Journal of Management Studies*, *57*(2), 351–376. [CrossRef]

Rethlefsen, M. L., Farrell, A. M., Osterhaus-Trzasko, L. C., & Brigham, T. J. (2015). Librarian co-authors correlated with higher quality reported search strategies in general internal medicine systematic reviews. *Journal of Clinical Epidemiology*, *68*(6), 617–626. [CrossRef]

Rizky Noviandy, T., Maulana, A., Mauer Idroes, G., Zahriah, Z., Paristiowati, M., Bin Emran, T., Ilyas, M., & Idroes, R. (2024). Embrace, don't avoid: Reimagining higher education with generative artificial intelligence. *Journal of Educational Management*, *2*(2), 81–90. [CrossRef]

Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). SAGE Publications Ltd.

Scott, A. M., Forbes, C., Clark, J., Carter, M., Glasziou, P., & Munn, Z. (2021). Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: A survey. *Journal of Clinical Epidemiology*, *138*, 80–94. [CrossRef] [PubMed]

Shin, D. (2025). *Debiasing AI: Rethinking the intersection of innovation and sustainability*. Routledge. [CrossRef]

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. [CrossRef]

Suanet, L. (2017). *Finding anomalies in sequential data using local outlier factor* [Bachelor's thesis, LIACS, Leiden University].

Sundelson, A. E., Jamison, A. M., Huhn, N., Pasquino, S.-L., & Sell, T. K. (2023). Fighting the infodemic: The 4 i framework for advancing communication and trust. *BMC Public Health*, *23*(1), 1662. [CrossRef]

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Pearson.

Tanimoto, T.-T. (1958). *An elementary mathematical theory of classification and prediction*. International Business Machines Corporation.

Temitope, O. A., Oyetola, O., & Folorunso Makinde, B. (2025). Artificial intelligence' receptiveness for inclusive education at Obafemi Awolowo University, Nigeria. *ETDC: Indonesian Journal of Research and Educational Review*, *5*, 431–443. [CrossRef]

VOSviewer. (2022). *VOSviewer—Visualizing scientific landscapes*. Available online: https://www.vosviewer.com/ (accessed on 24 January 2026).

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, *252*, 124167. [CrossRef]

Wang, W., Göpfert, T., & Stark, R. (2016). Data management in collaborative interdisciplinary research projects—Conclusions from the digitalization of research in sustainable manufacturing. *ISPRS International Journal of Geo-Information*, *5*(4), 41. [CrossRef]

Wätzold, F., Popiela, B., & Mayer, J. (2024). Methodology for AI-based search strategy of scientific papers: Exemplary search for hybrid and battery electric vehicles in the semantic scholar database. *Publications*, *12*(4), 49. [CrossRef]

Weaver, K. D. (2025). *AI tools for academic libraries: AI research assistants*. Available online: https://www.choice360.org/libtech-insight/ai-tools-for-academic-libraries-ai-research-assisants/ (accessed on 26 December 2025).

Wu, S., Ma, X., Luo, D., Li, L., Shi, X., Chang, X., Lin, X., Luo, R., Pei, C., Du, C., Zhao, Z.-J., & Gong, J. (2025). Automated literature research and review-generation method based on large language models. *National Science Review*, *12*(6), nwaf169. [CrossRef] [PubMed]

Zhan, Z., Shen, W., Xu, Z., Niu, S., & You, G. (2022). A bibliometric analysis of the global landscape on stem education (2004–2021): Towards global distribution, subject integration, and research trends. *Asia Pacific Journal of Innovation and Entrepreneurship*, *16*(2), 171–203. [CrossRef]