Review Method and Evidence Base

S1.1. Review Design

This Entry employed a sequential evidence-mapping and focused-synthesis design. The evidence-mapping stage identified the recent interdisciplinary literature on the evaluation and governance of artificial intelligence in education. The focused-synthesis stage applied narrower criteria to identify studies capable of informing the proposed auditing framework.

The review addressed two questions. First, what recent evidence exists on the evaluation and governance of AI systems used in educational settings? Second, what forms of evidence are required to determine whether an educational AI application is pedagogically defensible, developmentally appropriate, reliable, equitable, safe, contestable, and institutionally accountable? The design combined broad coverage of the literature with detailed examination of evidence directly relevant to these questions.

S1.2. Search scope and eligibility

Three sources provided complementary disciplinary coverage. Scopus covered research across education, computing, human-computer interaction, law, public administration, and related fields. ERIC provided additional coverage of education-specific scholarship. Crossref supported a supplementary search of eleven core education journals to identify relevant publications that differences in database indexing might otherwise have omitted. Supplementary File S2 reports the complete search strings, publication filters, journal titles and ISSNs, title-normalization procedures, and screening codes.

The review covered records published between 1 January 2024 and 30 June 2026. Scopus and Crossref records were filtered using day-level publication dates. Because ERIC metadata did not consistently provide complete day-level dates, ERIC records were filtered by publication year. The 2026 ERIC results therefore contain a small number of records published after 30 June. Supplementary File S2 documents this source-specific limitation.

Eligible records examined an AI system, its educational use, its users, or a documented governance mechanism in a formal, non-formal, or informal educational setting. The review considered empirical studies of tested or deployed systems, evidence syntheses relevant to educational AI evaluation, and analyses of existing policies, regulations, standards, or institutional governance arrangements. To qualify for focused synthesis, a record also had to address an identifiable audit object, evaluative procedure, educational consequence, or governance process relevant to at least one domain of the proposed framework.

The review excluded editorials, corrections, notices, bibliometric studies, purely conceptual proposals, and descriptions of systems that had not been evaluated. Studies concerned only with attitudes toward AI or intentions to adopt it were excluded unless they also presented evidence concerning system behavior, educational consequences, or governance. Records outside education and records that mentioned AI without examining an auditable system, outcome, risk, or governance mechanism were also excluded.

S1.3. Identification, deduplication, and eligibility screening

The searches identified 6,861 records: 3,245 from Scopus, 1,331 from ERIC, and 2,285 from Crossref. Application of the publication-period criterion retained all 3,245 Scopus records and 1,068 ERIC records. Application of the prespecified title-term filter retained 975 Crossref records. These filters produced 5,288 records, from which 163 duplicates were removed before screening, comprising one within-source duplicate and 162 cross-source duplicates.

The remaining 5,125 records underwent initial screening using their titles, abstracts, and available bibliographic metadata. This stage determined whether each record fell within the broad intersection of education, artificial intelligence, and system evaluation or governance. Screening excluded 4,136 records that did not meet these scope criteria, leaving 989 eligible records.

Cross-database duplication was assessed after the three source datasets had undergone initial screening. Titles were normalized to account for differences in capitalization, punctuation, spacing, and other minor variations in database records. This comparison identified eight additional duplicates across the three sources, leaving 981 unique records.

S1.4. Focused Screening and Full-Text Eligibility Assessment

The titles and abstracts of the 981 unique records were assessed against the focused eligibility criteria. Records advanced to full-text assessment when the available information indicated that they examined an identifiable AI application, audit object, evaluative procedure, educational outcome, risk, or governance mechanism relevant to the review questions. Records also needed to report sufficient methodological or analytical information to permit substantive evaluation. When the title and abstract did not provide enough information for a reliable decision, the record was retained for full-text assessment.

This stage excluded 912 records that met the broad scope of the evidence map but did not satisfy the requirements for focused synthesis. The principal reasons were the absence of an identifiable audit object or outcome, insufficient connection to the review questions, lack of evaluative evidence, or exclusive attention to AI adoption and perceptions without analysis of system behavior, educational consequences, or governance.

Sixty-nine records advanced to full-text eligibility assessment. Each was evaluated for the presence of evidence directly relevant to at least one review question and for sufficient information to determine the scope and basis of its findings. Thirty records were excluded because the full text did not provide relevant evaluative evidence, did not identify an audit procedure, object, or outcome, or duplicated evidence presented more completely in another eligible source. One primary reason was assigned to each excluded record to prevent double counting. The review ultimately included 39 records in the qualitative synthesis (Figure S1)

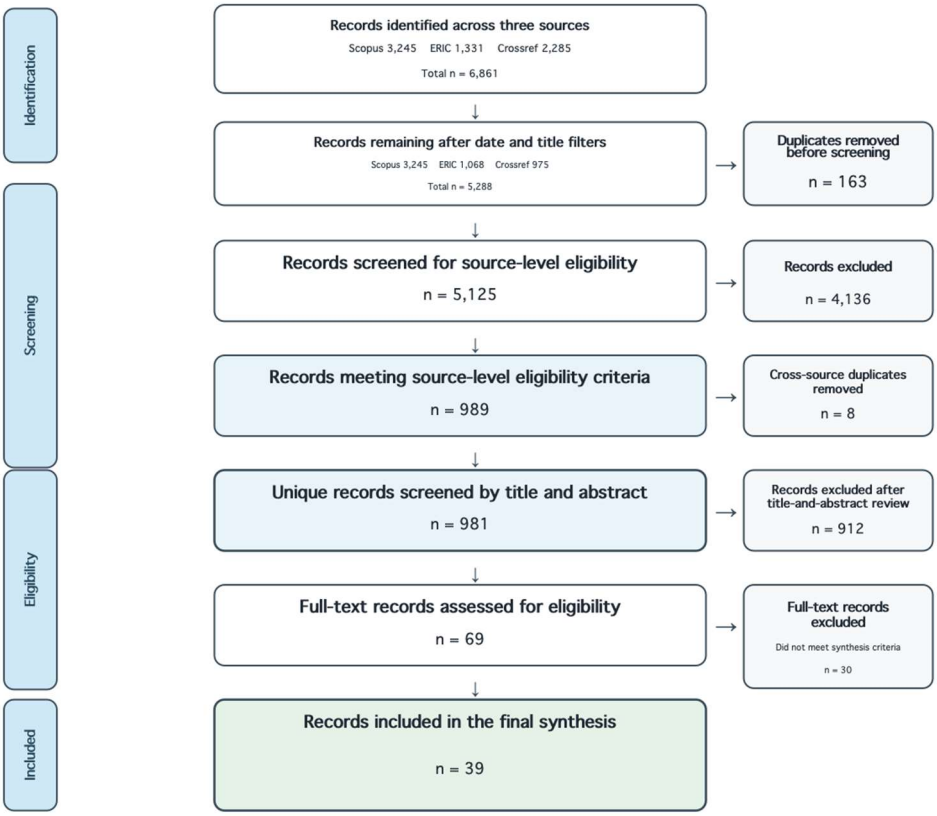

Figure S1. Identification, screening, eligibility assessment, and inclusion of records in the review.

S1.5. Data Extraction and Synthesis

Data were extracted from the 39 included records using a structured framework aligned with the review questions. For empirical sake, extraction covered the educational population and setting, AI function and deployment context, study design, comparison condition, outcome definition, reported findings, treatment of uncertainty, and stated limitations. For evidence syntheses, extraction addressed the scope of the review, characteristics of the underlying literature, synthesis method, and qualifications placed on the conclusions. For governance analyses, extraction documented the instrument or governance arrangement examined, the relevant jurisdiction or institutional setting, the analytical basis of the interpretation, and the scope of the resulting claims.

The heterogeneity of the included evidence precluded a meaningful pooled effect estimate. Studies differed in their populations, educational settings, AI functions, comparison conditions, outcome measures, and research designs. The review therefore used structured qualitative synthesis. Evidence was organized according to the object and level of the audit, the educational context, the type of evidence provided, and the domain of performance or risk addressed.

Conclusions were calibrated to the populations, settings, comparisons, and outcomes examined in the underlying sources. Convergent findings increased confidence in a conclusion, while contradictory, null, and conditional findings defined its boundaries. Methodological transparency, evidentiary directness, and contextual relevance informed the weight assigned to individual sources. No study was excluded because its findings conflicted with the proposed framework.

The 39 included records constitute the systematic review corpus. Foundational theoretical works and primary legal, policy, or professional documents cited solely to define concepts or establish applicable requirements serve a contextual or normative function. Such sources are identified separately and are not treated as additional records included in the systematic synthesis.

S1.6. Scope of inference and limitations

The findings support an analytical account of the evidence required to audit educational AI and the conditions under which conclusions about particular systems may be warranted. They do not support estimates of how frequently particular findings occur across the entire educational AI literature.

No meta-analysis was conducted because the included studies were not sufficiently comparable in their interventions, populations, comparison conditions, durations, outcome definitions, or analytical methods. A single conventional risk-of-bias instrument was not applied because the review incorporated heterogeneous empirical designs, evidence syntheses, and governance analyses. The interpretation of each source instead considered the transparency of its methods, the directness of its evidence, its contextual relevance, and its reported limitations. This approach supports claim-bounded qualitative synthesis but does not provide a standardized numerical estimate of study quality.

The review may also have been affected by the selected databases, the English-language restriction, variation in disciplinary terminology, publication lag, the limited publication window, differences in metadata quality, and the use of a single reviewer. The focused synthesis was restricted to evidence relevant to the specified review questions and should not be interpreted as an exhaustive account of every topic represented among the 981 unique records.

Much of the available evidence concerns adaptive learning systems or earlier generations of educational chatbots rather than current generative AI systems. Even where evidence syntheses report positive average effects, their estimates depend on the populations studied, comparison conditions, intervention durations, moderators, and outcome definitions. The Entry therefore does not infer that AI is generally effective or that an average result will transfer to a particular local deployment. Each conclusion is limited to what was tested, for whom, under what conditions, and with what degree of uncertainty. The Entry also distinguishes empirical findings from plausible risks, legal or ethical requirements, and recommendations advanced as part of the proposed framework.