A retrieval-augmented LLM framework for PRISMA-aligned systematic reviews: a case study on action quality assessment

Authors anonymized as requested[1*]

[1*]Affiliations anonymized as requested.

Abstract

Systematic reviews and meta-analyses are essential for advancing scientific research, but they are time-consuming, labor-intensive, and prone to human error during literature selection and data extraction. This paper proposes a framework that integrates Large Language Models (LLMs) and retrieval-augmented generation (RAG) into the PRISMA 2020 workflow to assist and standardize key review tasks. Our contribution is methodological rather than algorithmic: we operationalize its stages as a single reproducible, auditable pipeline built from established, openly available components, prioritizing transparency and re-execution over customized elements. The framework uses structured query templates to ensure reproducible bibliographic searches, an LLM-driven pipeline for title and abstract screening, and a RAG-based full-text eligibility assessment. Both automated stages rely on structured prompt templates that encode inclusion and exclusion criteria as machine-interpretable instructions. Beyond selecting relevant papers, it extracts specific information from documents, such as application details or datasets, and harmonizes it into a standardized metadata schema. To evaluate our proposal, we performed an experimental validation using an existing systematic literature review on Human Action Quality Assessment (AQA) as a reference. The framework achieved an extraction recall of 85.11% for domain-specific datasets and identified 25 additional data sources not reported in the original review. The pipeline achieved an overall article detection rate of 71.28%. This reduction was primarily a consequence of the strict, dataset-oriented Boolean query adopted in the identification phase, rather than failures in the automated LLM or RAG components. Supplementary analyses confirmed near-deterministic behavior across repeated runs and successful transfer to a different LLM family at negligible computational cost. These results suggest that the framework can effectively assist in identifying relevant studies and extracting information while maintaining alignment with PRISMA 2020 guidelines. The

proposed approach is structured to facilitate adaptation to other research areas by declaring new query and prompt templates.

Keywords: PRISMA 2020, systematic review, large language models, retrieval-augmented generation

1 Introduction

Systematic reviews are essential to scientific research because they synthesize evidence in a structured way, support state-of-the-art assessment, and help identify current limitations and research gaps [5]. Yet, conducting them remains labor-intensive, often exceeding 1,000 person-hours due to extensive literature searches with low inclusion rates [7]. Furthermore, this process is typically sensitive to human variability, with screening errors ranging from 10% to 20% [41], and worst-case values reaching almost 40% [51]. Early pioneering work on automation in research synthesis showed that machine learning can support this process [35, 49], but many current solutions still address isolated steps (*e.g.*, screening) rather than the full workflow, or do not align with reporting standards such as PRISMA [37].

Large Language Models (LLMs), *i.e.*, transformer-based models trained on large text corpora, are increasingly used in systematic-review support tasks [53]. Khraisha et al. [31] showed that GPT-4 [1] can achieve accuracy above 80% in full-text screening on imbalanced datasets with carefully designed prompts, a specific setting that may not generalize to other screening configurations or model families.

However, many current approaches remain fragmented and do not operationalize the PRISMA workflow end-to-end. Existing studies also show limited workflow-level integration of retrieval-augmented generation (RAG) across PRISMA stages [15, 34, 47], which can reduce traceability and auditability of stage-wise decisions [44]. Although RAG improves factual grounding at the component level [21], its methodological integration in PRISMA-aligned end-to-end workflows remains limited.

In this paper, we present a unified framework that incorporates structured literature searches, LLM-driven screening, and RAG-based full-text assessment within the PRISMA 2020 workflow, an updated guideline that refines the original PRISMA [40]. We treat PRISMA as an operational framework to enforce transparent reporting, traceable decisions, and reproducible procedures across the full pipeline. Our contribution is methodological rather than algorithmic. Built entirely from established, openly available components, its novelty lies in mapping the PRISMA 2020 stages onto a single reproducible, auditable pipeline, embedding structured extraction within screening and eligibility decisions, and providing artifacts that make each stage re-executable. This design lowers the barrier to others reproducing and rerunning the workflow.

In addition to identifying relevant studies, the framework extracts domain-specific information, such as datasets and application details, and organizes it into a consistent metadata schema. By integrating LLMs and RAG into a methodologically grounded PRISMA framework, our approach aims to support systematic reviews by reducing

manual effort and to provide a practical tool for transparent and reproducible research synthesis.

To evaluate the proposed framework, we conducted an empirical evaluation in one target domain, Human Action Quality Assessment (AQA). First, we selected a manually curated, peer-reviewed reference study [33] as the comparison baseline and replicated its bibliographic identification procedure as closely as possible. Then we applied our automated pipeline under comparable conditions. The evaluation focused on methodological reproducibility and on the framework's ability to identify relevant studies and extract related information, not on computational performance.

We assessed how closely the resulting corpus and extracted outputs matched the reference baseline in terms of coverage and relevance. The framework achieved a reference-study detection rate of 71.28% and a reference-dataset extraction recall of 85.11%. Most discrepancies originated in the initial bibliographic search, while the automated screening and eligibility stages accounted for only a marginal loss, retaining 95.71% of the reference studies available at the pipeline entry. This evaluation is a single-domain, single-reference case study with one primary model configuration and one supplementary cross-model execution; the results therefore establish feasibility in this setting rather than general validity.

Although validated on AQA in this study, the workflow is designed to allow adaptation across domains. By re-specifying query terms and extraction criteria in the prompt templates, the same PRISMA-aligned framework could potentially be applied to other research areas, without modifying the pipeline itself. However, this transferability remains to be empirically verified in future work.

The main contributions of this work are as follows:

- We propose a methodological operationalization of the PRISMA 2020 workflow as an end-to-end, reproducible pipeline that integrates structured query design, LLM-based screening, and RAG-based full-text eligibility assessment, assembled from established, openly available components to prioritize transparency and re-execution.
- We formalize operational artifacts for reproducibility, including query templates, prompt templates, harmonization rules, and machine-readable outputs for each PRISMA stage.
- We provide a single-domain empirical validation against a manually curated reference study, demonstrating substantial agreement in study retrieval and data-source extraction, and we characterize robustness of these results in terms of run-to-run stability across five independent runs, computational cost, and portability across LLM families.
- We demonstrate that the framework can identify additional relevant data sources beyond those reported in the reference study, supporting its utility for evidence discovery in evolving research domains.

The remainder of this work is organized as follows. Section 2 discusses the relevant literature. Section 3 presents our proposed methodology, while Section 4 describes the experimental validation in a case-study setting, focusing on reproducibility and the ability to identify relevant studies and information in a manner comparable to a

traditional manual review. This section also presents the reference study, its workflow and results, our partial replication of its approach, and the technical implementation of our framework. Section 5 discusses the evaluation results, including robustness and portability analyses, and Section 6 concludes the paper with final remarks.

2 Related work

Systematizing the synthesis of scientific evidence has become a critical requirement as the volume of global literature grows exponentially [5]. The PRISMA framework was originally introduced to standardize this process [37] and subsequently updated in PRISMA 2020 to incorporate advances in existing methods and technologies [40]. Extensions such as PRISMA-S [42] have been developed to provide detailed guidance on reporting search strategies across multiple databases, supporting transparency and reproducibility. Yet, the practical application of these protocols still relies heavily on manual workflows, which are time-consuming and prone to human error, motivating research interest in automated solutions. Recent broad mappings of the field indicate that evidence-synthesis automation now spans both classical machine learning tools and LLM-based systems; however, the landscape remains heterogeneous in terms of workflow coverage, reporting practices, and validation depth [25, 30].

Early efforts in this domain focused on semi-automated title/abstract screening and structured information extraction to support evidence synthesis. Tools such as RobotReviewer applied NLP to assist with risk-of-bias assessment in randomized controlled trials [35]. In parallel, active-learning-based screening systems (*e.g.*, ASReview [49], Rayyan [39]) implemented a reviewer-in-the-loop workflow in which the model is retrained after each reviewer label and used to rank the remaining records for manual screening, yielding substantial reductions in screening effort at high target recall. In this context, a complementary study by de la Torre-López et al. [48] describes automated pipelines for classification, clustering, and extraction based on discriminative models. Collectively, these approaches reinforced that automation in systematic reviewing must remain auditable and conservative, with clear traceability of decisions and outputs [44]. Khalil et al. [30] offered a scoping overview of available automation tools, their validation status, and limitations. More recent reviews confirm that, although the field is expanding rapidly, many solutions remain task-specific rather than fully end-to-end [13, 25].

Recent advances in LLM research extend earlier discriminative approaches; unlike prior tools focused on narrow classification, LLMs introduce generative capabilities and richer semantic understanding to the review process [53]. These technologies have paved the way for systems that integrate diverse data sources to support researchers during the selection and extraction phases [49]. At the same time, these studies highlight the variability of LLM outputs, stressing the need for validation mechanisms and traceable reporting of model decisions [38]. Empirical evaluations of LLM-assisted workflows further illustrate this trade-off. Scherbakov et al. [43] test an LLM-based systematic review workflow and report notable efficiency gains paired with the continued need for human oversight to manage errors. Broader analyses also discuss ethical and transparency concerns around assisted reviews [6, 50]. Recent studies have

examined how LLMs can operationalize PRISMA principles. Thode et al. [47] and Delgado-Chaves et al. [15] studied the role of LLMs in the screening phase and reported significant increases in efficiency. Their empirical evaluations suggest that although LLMs can substantially reduce workload, human supervision remains essential to mitigate risks of hallucinations or misinterpretations of text. Complementarily, Galli et al. [20] examined the use of LLMs in abstract screening, highlighting prompt engineering, zero-shot/few-shot classification, and the scalability constraints of AI-assisted screening. More specifically, Huotala et al. [27] compared GPT-3.5 and GPT-4 under zero-shot, one-shot, few-shot, and few-shot chain-of-thought prompting for title-and-abstract screening, while Cao et al. [11] developed and evaluated adaptable prompt templates for abstract and full-text screening across ten systematic reviews and compared multiple GPT-4 variants with Claude 3.5. To formalize this supervision, recent research has proposed hybrid methodological frameworks: Brincoveanu et al. [8] developed a collaborative human-AI environment to minimize error rates, while Malik and Terzidis [34] outlined specific operational guidelines for augmenting PRISMA workflows with AI checkpoints. Furthermore, RAG has emerged as a key technical method for grounding LLM outputs in source documents, thereby improving factuality [21, 32]. Recent surveys map the broader RAG design space, but also show that practical integration into evidence-synthesis workflows and explicit reasoning support remain limited [9, 36].

Regarding reporting standards, the emergence of Generative AI (GenAI) has highlighted the need to adapt guidelines to AI-assisted review workflows. Recent governance-oriented contributions, including the Cochrane statement and the PRISMA-trAIce checklist, emphasize explicit reporting of AI-specific design choices [18, 26]. For example, Shailendra et al. [45] present L-PRISMA, a preprint that proposes reporting extensions for AI-specific elements such as model provenance, prompt design, and verification procedures. In parallel, Forero et al. [19] applied LLMs to assess the adherence of published reviews to PRISMA 2020 items, showing the potential of these models as auditing tools. Finally, toward broader automation, agent-based pipelines have been proposed to automate multiple stages of the review process. Wang et al. [52] proposed TrialMind, an LLM pipeline for clinical evidence synthesis integrating study search, citation screening, data extraction, and evidence synthesis, evaluated on a dedicated benchmark of 100 published systematic reviews. Notably, TrialMind employs RAG specifically within its query refinement step during literature search, using retrieved study abstracts to enrich and expand Boolean queries, whereas the framework proposed here applies RAG to full-text eligibility assessment, a methodologically distinct application within the PRISMA workflow. Cao et al. [10] introduced a multi-agent system capable of executing sequential tasks, and Susnjak [46] explored the use of fine-tuned models for specific domains (PRISMA-DFILM), with improved relevance in targeted application settings.

To systematically compare these approaches, we evaluate them across four operational dimensions using a defined rubric. **Reproducibility** assesses the availability of replication artifacts: *High* (full declaration of models, prompts, parameters, and decision logic, with codebase accessible to re-execute the end-to-end pipeline), *Moderate* (most operational parameters reported, such as prompts or model configuration,

but insufficient artifacts for end-to-end re-execution), *Limited* (methodology described in narrative form, but artifacts insufficient for re-execution), or *None* (no reusable artifacts reported). **Auditability** evaluates the transparency of automated decisions: *High* (decision-level provenance for every stage, including explicit logging of intermediate outputs and fallback routing), *Moderate* (retention of intermediate outputs for a subset of stages, without formal fallback logging), *Limited* (aggregate statistics only, with no decision-level traceability), or *None* (black-box pipeline with no traceability). **Attribute Extraction** measures data synthesis capabilities: *High* (automated extraction and harmonization of domain-specific entities), *Moderate* (basic metadata extraction without harmonization), *Limited* (manual extraction only), or *None* (classification only, no extraction). Finally, **RAG use** indicates the integration of an explicit retrieval-augmented generation mechanism, defined here as a pipeline combining a persistent vector store with generation grounded in retrieved chunks, as distinguished from retrieval used only for ranking or filtering.

Taken together, these studies show substantial progress in automating individual review tasks. Table 1 summarizes representative approaches according to PRISMA-stage coverage, integration level, use of RAG, and reproducibility and auditability criteria. Commercial and proprietary platforms, such as Elicit, Covidence, and DistillerSR[1], were not included in the structured comparison because the publicly available documentation does not provide sufficient information to apply the reproducibility and auditability rubric consistently and without unsupported assumptions. This exclusion concerns only the applicability of the selected comparison criteria and should not be interpreted as an assessment of their practical relevance, maturity, or effectiveness. Among the publicly documented research systems included in the comparison, the landscape remains largely characterized by stage-specific or partially integrated solutions, with limited simultaneous coverage of PRISMA alignment, explicit reproducibility artifacts, and retrieval-grounded processing across multiple workflow stages. This observation motivates the workflow-level contribution presented in this study.

[1] https://elicit.com/, https://www.covidence.org/, https://www.distillersr.com/

Table 1 Comparison of representative AI-assisted systematic review approaches with respect to PRISMA workflow coverage and methodological rigor.

Study	PRISMA stage(s)	Approach type	Reproducibility	Auditability	Attribute Extraction	RAG use
[35]	Inclusion	Task-specific Tool	Limited	Moderate	Moderate	No
[49]	Screening	Task-specific Tool	Moderate	Moderate	None	No
[39]	Screening	Task-specific Tool	Limited	Moderate	None	No
[15, 47]	Screening	Partial LLM Workflow	Limited	Moderate	None	No
[34, 45]	All stages	Theoretical Framework	None	High	None	No
[48]	Screening, Inclusion	Partial Workflow	Limited	Moderate	Moderate	No
[43]	Screening, Eligibility, Inclusion	Partial LLM Workflow	Moderate	Moderate	High	No
[20]	Screening	Partial LLM Workflow	Limited	Moderate	None	No
[8]	Screening, Eligibility	Partial LLM Workflow	Moderate	High	Moderate	No
[19]	N/A	Audit Tool	Moderate	High	None	No
[52]	All stages	End-to-end LLM Workflow	High	High	High	Yes
[10]	Screening, Eligibility, Inclusion	Multi-agent LLM Framework	Limited	Moderate	High	No
[46]	Screening, Eligibility	Task-specific Tool	Moderate	Moderate	Moderate	No
This work	**All stages**	**End-to-end LLM Framework**	**High**	**High**	**High**	**Yes**

Note: Reproducibility, Auditability, and Attribute Extraction are rated according to the four-level rubric defined above. RAG use indicates the integration of an explicit retrieval-augmented generation mechanism, as defined above. N/A denotes tools not directly mapped to a PRISMA stage.

3 Proposed methodology

In this section, we describe our proposal for a methodology that extends the PRISMA 2020 workflow for systematic reviews [40] to integrate LLM and RAG components. Our workflow, shown in Figure 1, follows the original four phases of the PRISMA

2020 methodology (*i.e.*, identification, screening, eligibility, and inclusion) while integrating automated and semi-automated components. Each automated component is associated with a set of parameters and outputs that must be reported to ensure reproducibility. In conventional PRISMA-aligned reviews, data extraction is formally prescribed as a post-selection step, conducted once the final corpus has been established. In the proposed framework, attribute extraction is instead embedded directly within the screening and eligibility phases, allowing extracted information to propagate across pipeline stages and reducing redundant effort on records that will ultimately be excluded. This embedding departs from PRISMA 2020's conceptual separation between selection and extraction. In this case study, the risk of bias is mitigated because dataset mention is itself one of the inclusion criteria (Figure 7): extraction and eligibility assessment are therefore aligned by construction. This alignment, however, is specific to dataset-oriented review objectives; in domains where extracted attributes are unrelated to the inclusion criteria, the conceptual separation prescribed by PRISMA 2020 should be preserved. In the following, we provide a detailed description of each phase, focusing on the theoretical and procedural alignment with PRISMA 2020 requirements and highlighting the novel procedural elements introduced by this framework.

3.1 Identification: information sources and search strategy

The identification phase aims to design and execute search queries across multiple bibliographic search engines. Queries are composed of a fixed set of declarative components that together delimit the scope of the search and standardize the query formulation process. For each search, the exact query string should be declared along with the date on which it was executed and the bibliographic search engine used. Declaring these details follows the PRISMA-S recommendations for reporting complete search strategies [42], ensuring transparency and supporting reproducibility.

Conceptually, we structure a systematic search query around four core elements:

- **Domain / Subdomain**: the principal research area and any narrower subfields, guiding the domain-specific terminology used in the query;
- **Object of study**: the phenomenon, method, or element of interest that drives core topical terms of the search;
- **Period of reference**: explicit temporal bounds (publication year range) used to control temporal bias and support reproducible results;
- **Qualifiers**: optional filters such as publication type, geographic focus, language, or other attributes that restrict the search according to the review objectives.

To operationalize this conceptual structure, we utilize parameterized search strings, defined here as query templates. A query template is a reusable structure composed of named fields corresponding to the four components described above, together with angle-bracketed placeholders (*i.e.*, strings intended to be replaced by specific terms, *e.g.*, `<DOMAIN_TERMS>`) that are instantiated with keywords and Boolean operators. While PRISMA-S prescribes the reporting requirements for search strategies, it does not provide operational structures for constructing them. We present these templates

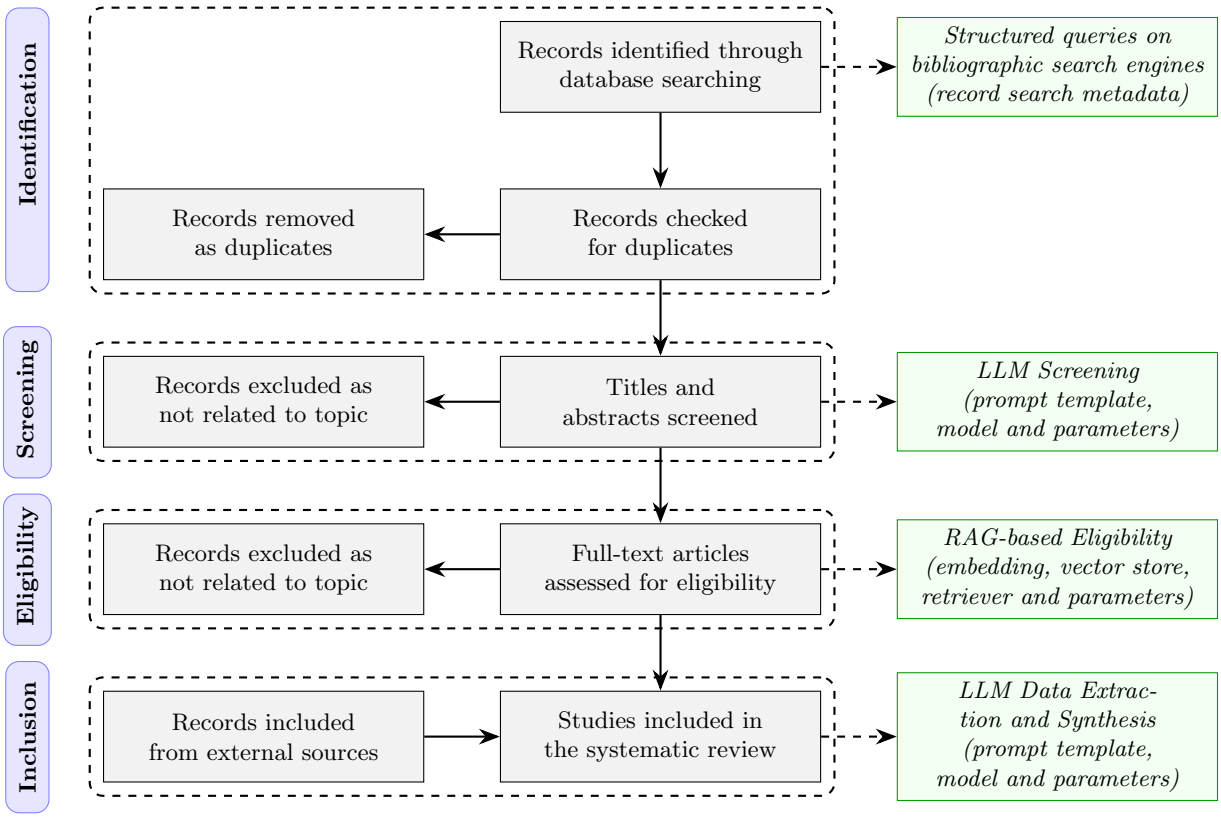

Fig. 1 Proposed workflow, with integrated LLM and RAG components, in reference to the traditional PRISMA 2020 diagram [40].

as a practical engineering mechanism that bridges this gap, providing a reusable operational structure for consistent query construction across databases. Templates also encode the syntax rules of the target bibliographic search engine and preserve the logical structure of the query. Each field in a template is populated with all relevant terms corresponding to its component, including synonyms and acronyms, to represent the domain or object of study. From an applied perspective, the implementation of query templates offers four practical benefits. First, they provide a structured framework that is reusable and adaptable to future systematic reviews on related topics. Second, they organize queries into distinct components, thereby clearly delimiting the research objective at this stage of the workflow. Third, they increase robustness to lexical variation by systematically covering both domain-specific and object-specific terminology. Fourth, they ensure transparency and reproducibility by making the exact executed queries explicit and auditable.

As an example, a topic-agnostic Scopus query template is presented in Figure 2, illustrating the query structure without specifying a particular research topic. The

```
TITLE-ABS-KEY ( <DOMAIN_TERMS> )
AND TITLE-ABS-KEY ( <OBJECT_TERMS> )
AND PUBYEAR > <START_YEAR> AND PUBYEAR < <END_YEAR>
AND ( LIMIT-TO ( DOCTYPE , <DOCTYPE_FILTERS> ) )
AND ( LIMIT-TO ( SUBJAREA , <SUBJAREA_FILTERS> ) )
AND ( LIMIT-TO ( LANGUAGE , <LANGUAGE_FILTERS> ) )
```

Fig. 2 Topic-agnostic Scopus search template for systematic review searches.

angle-bracketed placeholders are replaced with the terms specific to the review under consideration, according to the conceptual structure defined above.

Executing queries on search engines returns a list of documents (*i.e.*, papers) stored in a centralized repository. Each document record is normalized to a structured metadata schema to ensure consistency across sources and support data processing. The schema presented in Table 2 is a practical metadata model proposed by the authors to capture key bibliographic details of each record, including the identifier, title, publication year, abstract, source, and other optional attributes. While PRISMA 2020 recommends reporting the number of retrieved records and the search strategies adopted, it does not prescribe a specific metadata schema. The schema proposed here aligns with standard bibliographic metadata conventions and provides a reproducible basis for the subsequent phases.

As a final step in the identification phase, deduplication is required by PRISMA 2020. The goal is to remove all redundant copies of the same record. When persistent identifiers (*e.g.*, DOI or PMID) are available, simple deterministic rules are applied to remove duplicates. For records lacking reliable identifiers, which may occur when integrating results from grey literature or non-standard sources, reproducible fuzzy-matching heuristics may be applied, or records may be resolved through manual assessment.

3.2 Screening: title and abstract evaluation

The screening phase performs an initial eligibility assessment at the title and abstract level using an LLM-driven classification pipeline (hereafter referred to as the "screening pipeline"), as illustrated in Figure 3. This stage corresponds to the PRISMA screening phase and is designed to reduce the candidate set while preserving alignment between the article content and the predefined inclusion and exclusion criteria for the systematic review. The screening stage operates exclusively on bibliographic metadata generated in the identification stage and does not require access to full-text content.

Methodologically, the screening pipeline is organized into two main conceptual steps:

1. Ingestion: Bibliographic records are prepared for LLM evaluation by extracting the title and abstract of each study and grouping them into batches. A batch is a fixed-size subset of records that are jointly submitted to the LLM in a single prompt. Batch size is a parameter that must be explicitly declared: smaller batches improve output consistency but increase processing time and API calls, while larger batches improve efficiency but may degrade screening quality if the cumulative

Table 2 Metadata schema required for each record in the identification phase.

Metadata field	Description	Purpose	Required
Identifier	Persistent and unique identifier (*e.g.*, DOI, PMID)	Ensure unambiguous identification and traceability across all systematic review stages	Yes
Title	Full record title	Support initial eligibility assessment in the PRISMA screening phase	Yes
Publication year	Year of publication	Enable verification of reference period and reproducibility of the review	Yes
Abstract	Study summary	Support PRISMA screening and subsequent data extraction	Yes
Source	Bibliographic source or search engine	Verify provenance and ensure traceability of the resource	Yes
Authors	List of authors as reported by the source	Support provenance tracking and optional verification during deduplication	Optional
Keywords	Author or indexer assigned keywords	Assist eligibility filtering and verification in later stages	Optional
Document type	Type of document (*e.g.*, article, conference paper, review)	Assist eligibility filtering and verification in later stages	Optional
Language	Language of publication	Assist eligibility filtering and verification in later stages	Optional

input approaches the LLM's context window limit. The batch size must therefore be chosen in relation to the context window of the selected model and reported as part of the screening metadata.

2. Generation: Using the batches prepared in the ingestion step, each group of records is evaluated by providing the LLM with the title and abstract, along with a prompt template that encodes the inclusion and exclusion criteria as machine-interpretable instructions. Each batch is dynamically concatenated into the prompt at runtime, ensuring that the template itself remains fixed. The prompt is applied uniformly across all records to ensure reproducibility and methodological consistency. The LLM is required to return a structured response that conforms to the *Output schema* field specified in Table 3. Following LLM evaluation, responses are automatically validated against the schema (*i.e.*, output validation). Finally, selected attributes are standardized in a subsequent harmonization step to ensure consistency across records.

A key difference between this workflow and manual screening is the shift from subjective human assessment to an automated process. Upon executing these steps, the pipeline produces a structured object for each screened record. This output includes:

- a categorical screening decision (*include*, *exclude*, or *uncertain*);
- a short justification that links the decision to the stated eligibility criteria;
- any extracted study-level fields required for subsequent analyses (*e.g.*, named data sources or other identifying attributes).

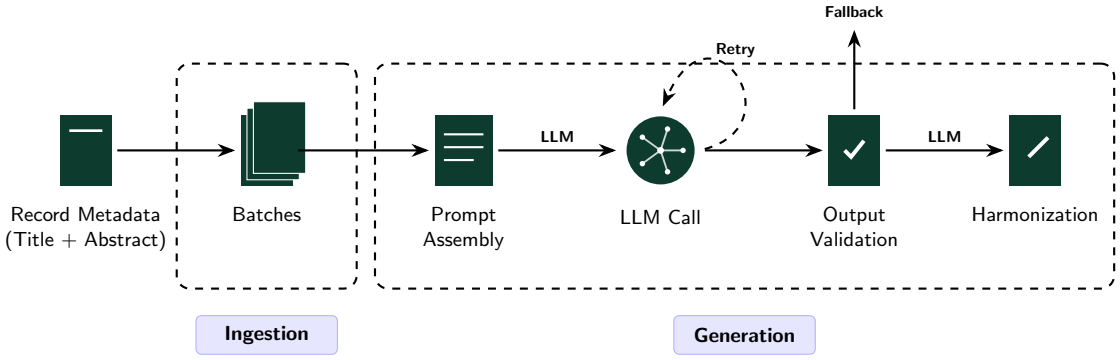

Fig. 3 Screening pipeline supporting title and abstract evaluation under PRISMA workflow.

To ensure the reliability of this automated process, all LLM calls use model parameters configured to minimize output variability, and responses are validated against the expected *Output schema* prior to acceptance. Outputs that fail schema validation, are incomplete, or cannot be reliably parsed are not forcefully classified. Instead, such cases are conservatively labeled as *uncertain* and either passed to subsequent workflow steps or flagged for human review. A conservative fallback policy is therefore intrinsic to the screening design: when automated classification is not completed with sufficient confidence, records are kept rather than discarded. This approach minimizes the risk of false negatives during the screening stage. In this context, the justification accompanying each decision serves primarily as a verification aid for the human reviewer during these manual checks, rather than as the basis for the categorical decision itself.

In addition to the screening decision, selected attributes can be extracted from the LLM output for use in later phases of the workflow. These fields undergo a dedicated harmonization step to reduce variability in attribute names, such as differences in capitalization, punctuation, abbreviations, or descriptive modifiers. Typical normalization rules include lowercasing, trimming whitespace, removing diacritics such as accent marks and similar signs (*e.g.*, "Intérnational" to "International"), standardizing punctuation, and expanding common abbreviations (*e.g.*, "ML" to "Machine Learning").

Robustness measures are integrated to handle model errors and execution failures. The screening pipeline includes retry and error-management policies to prevent data loss during temporary service interruptions. Human supervision is explicitly integrated as a complementary component of the screening phase. First, manual review is reserved for all records labeled *uncertain*. Second, a configurable sample of automatically accepted and discarded records, typically in the range from 5% to 10% of each group, is examined for quality control purposes. Finally, manual review is conducted for cases where the automated harmonization results are ambiguous. The frequency and scope of manual checks are recorded and may be used to refine screening criteria or calibration strategies.

Table 3 Operational artifacts and metadata preserved for the LLM-assisted title and abstract screening phase.

Item	Description	Reporting Purpose	Required
LLM Specifics	Identifier, version and parameters of the language model	Ensure LLM provenance and enable comparison of automated decisions across runs	Yes
Prompt template	Instruction template encoding the inclusion and exclusion criteria for screening	Support reproducibility of eligibility rules and related instructions	Yes
Output schema	Machine-readable response format (*e.g.*, JSON Schema or Pydantic)	Ensure consistent parsing and alignment of expected outputs	Yes
Fallback policy	Rules for fallback labels, recovery, and manual review routing	Specify handling of non-conforming outputs and possible escalation to human review	Yes
Batching strategy	Batch size and grouping logic for processing	Support efficient processing and consistency of results	Optional
Harmonization rules	Normalization rules for extracted fields	Document standardization and mapping of extracted fields	Optional
Provenance logging	Metadata recorded for screening decisions at record level	Enable traceability of decisions in the screening phase	Optional

Table 3 summarizes the key items reported from the use of an LLM-based screening procedure at the title and abstract level. Most items capture metadata about the phase as a whole, while provenance logging is recorded for each individual record.

3.3 Eligibility: full-text assessment via RAG

The automated eligibility phase implements the full-text assessment under PRISMA 2020 requirements using a RAG-based pipeline. Unlike the standard PRISMA methodology, which assumes manual reading of full-text documents by the researchers involved, this framework introduces automated retrieval as the primary mechanism for evidence selection. Full-text articles are first acquired from their original sources and stored in a repository. Each document is then transformed into a numerical vector representation of text (*i.e.*, embeddings) using an embedding model, and these embeddings are indexed in a persistent vector store (*i.e.*, a database optimized for searching information) to enable efficient semantic similarity search. During assessment, relevant context windows are retrieved from the vector store and provided as input to the LLM. The choice of embedding model is detailed in Section 4.3, while the retrieval configuration and similarity search specifications are described in Section 4.4.3.

Methodologically, the eligibility pipeline is organized into three conceptual steps, as illustrated in Figure 4:

1. Representation: Full-text articles are segmented into retrievable units, called chunks (*e.g.*, logical sections, paragraphs, or fixed-length segments), and encoded into dense vector embeddings. Embeddings are stored in a vector store alongside basic provenance metadata, such as the document identifier and chunk origin, enabling robust semantic retrieval that is insensitive to lexical variation.

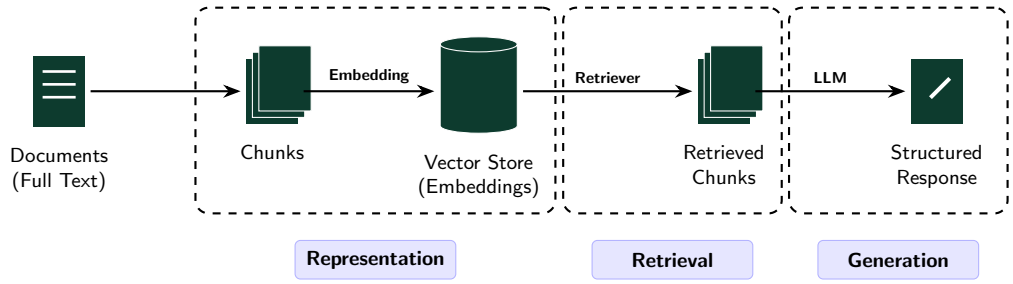

Fig. 4 Eligibility pipeline supporting full-text eligibility assessment under PRISMA workflow. For the sake of space, the Generation phase is shown here in a simplified form; its detailed architecture is provided in Figure 3.

2. Retrieval: Using the vector store populated in the previous step, a similarity-based retriever identifies the most relevant information using a retrieval query (*i.e.*, a natural language string or set of keywords derived from the inclusion and exclusion criteria). For this query, the retriever selects a bounded set of candidate chunks per document that are most likely to contain information relevant to the prespecified inclusion and exclusion criteria. Retrieval outputs include ranked relevance scores and explicit provenance pointers.
3. Generation: Based on the candidate chunks provided by the retriever, the LLM analyzes a prompt template that integrates the chunks with the inclusion and exclusion criteria. The model then produces structured responses that can be used to assign a categorical eligibility decision at the full-text level and to extract standardized study attributes that are useful for the review.

During execution of the eligibility pipeline, several constraints are enforced to maintain rigor and mitigate LLM limitations. LLM outputs are required to conform to a predefined schema (*e.g.*, a JSON schema specifying fields for decision and extracted attributes). Consistent with the screening phase, outputs are automatically validated against this schema, and any response that fails validation is flagged and routed to human review in the following inclusion phase rather than automatically accepted. All operational parameters that could influence LLM behavior must be declared to ensure reproducibility. In the proposed framework, these parameters are reported as part of the workflow metadata; their specific values are detailed in Section 4.3.

The same robustness measures as in Section 3.2 are integrated at key steps of the pipeline — including retry policies, rate limiting, output validation, and conservative fallback routing — with the addition of document-level provenance tracking, which records the origin of each retrieved chunk at query time.

All reporting items listed in Table 3 also apply to the full-text eligibility assessment. Table 4 further specifies the additional items that govern the steps of the eligibility pipeline, as they define how full-text documents are acquired, segmented, embedded, and retrieved to provide the evidence base for automated eligibility decisions.

It is important to emphasize that the reporting schemas presented in Table 3 and Table 4 are not proposed as novel alternatives to established reporting standards,

Table 4 Additional operational artifacts and metadata preserved for RAG-based full-text eligibility assessment.

Item	Description	Reporting Purpose	Required
Embedding Specifics	Identifier, version, and parameters of the embedding model	Enable reproducibility of semantic representations and consistency of retrieval behavior	Yes
Document acquisition	Procedure for obtaining the full texts	Ensure provenance and traceability of source documents	Yes
Segmentation strategy	Method for dividing documents into chunks	Support appropriate retrieval granularity and contextual coverage	Yes
Retriever configuration	Set of parameters such as similarity metric and other filtering rules	Specify how candidate evidence chunks are selected for generation	Yes

such as PRISMA-trAIce [26] or L-PRISMA [45]. Rather, they serve as machine-level operational logs. While guidelines like PRISMA-trAIce dictate what must be disclosed to ensure transparency, our schemas define the underlying data structures that the framework captures during execution.

3.4 Inclusion: final selection and reporting

The inclusion phase is the final step of the proposed workflow and corresponds directly to the inclusion phase as defined in PRISMA 2020 [40]. It consolidates the set of studies that have passed all screening and eligibility assessments and prepares them for data synthesis.

In the proposed framework, this phase takes as input the records that were assessed as eligible in the previous phase (*i.e.*, those for which the RAG-based pipeline produced a complete, validated, and harmonized structured response) and marks them as included in the systematic review. Records for which extraction was incomplete or schema validation failed are not automatically excluded; instead, they are flagged for manual verification and curation, in line with PRISMA 2020's requirement for transparent reporting of the selection process and human oversight of uncertain cases.

In addition to confirming eligibility, the inclusion phase performs supplementary functions. First, it supports the integration of records from external sources, such as domain knowledge or complementary searches, provided that these records undergo the same screening and eligibility pipeline as the primary corpus. This is consistent with the PRISMA 2020 guidelines for reporting additional sources. Second, it performs a final manual validation of the attributes extracted incrementally during the screening and eligibility phases (*e.g.*, study identifiers, methods, results, or domain-specific fields) to ensure their correctness before synthesis. Third, it maintains full traceability for each included study, linking the original bibliographic metadata, full-text sources, retrieved chunks, and generative outputs into a complete provenance trail.

Therefore, the proposed framework does not alter the scope or criteria of the PRISMA 2020 inclusion phase, but operationalizes them through structured information extraction. This approach is a procedural refinement that supports automated processing while preserving compliance with reporting requirements.

4 Experimental validation

To evaluate the quality of our proposal, we conducted an experimental validation within a specific research domain. This validation emphasizes the reproducibility of the proposed approach rather than the performance of its software implementation. The methodology aims to identify relevant papers and related datasets (as an example of specific information) in alignment with a traditional, manually curated process. The expected outcome need not exactly match manual review decisions; instead, it should achieve a comparable level of coverage and relevance in the resulting collection of studies and associated data.

In the remainder of this section, we present the reference study described in [33], including its workflow and results. Additionally, we replicate the methodology described by the authors in the original work to the best of our ability, using all available information from the study, and report those results as well. We then describe the technical implementation of our methodology and, finally, how we applied it to the same domain as the reference study. The results of this final step, along with their comparison to the reference study, are presented in Section 5.

4.1 Reference case study

As a reference, we adopted the systematic review by Liu et al. [33]. This review focuses on AQA and examines aspects such as existing datasets, applications, equipment, models, and evaluation metrics. We selected this study because it provides a clearly defined object of interest, explicitly adheres to the PRISMA methodology, employs multiple established bibliographic databases, and provides sufficiently detailed metadata to enable a meaningful comparison. Additionally, it is a recent peer-reviewed study, ensuring alignment with current literature and terminology. From a difficulty standpoint, AQA represents a moderately specialized domain with partially standardized terminology, making exhaustive retrieval non-trivial and therefore a meaningful test case for the proposed framework. For our evaluation, we focused on extracting and analyzing existing datasets as additional domain data. To establish a rigorous evaluation baseline, we performed a comprehensive manual audit of all original records and datasets compiled in the reference systematic review. This preliminary screening allowed us to manually verify the validity, taxonomy, and contextual relevance of each referenced dataset, ensuring that the human baseline against which the pipeline is compared constitutes a reliable benchmark. In the following paragraphs, we provide the details relevant to comparing our methodology. For further information, the reader is referred to the original paper.

The reference review relied on three major bibliographic databases: IEEE Xplore, Scopus, and Web of Science. Figure 5 reports the search query template adopted in the study, using Scopus as a representative example. Equivalent queries, adapted to

```
TITLE-ABS-KEY ( "action quality assessment" OR "action
    assessment" OR "action quality evaluation" OR "human
    action evaluation" OR "movement quality assessment" )
AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO ( DOCTYPE , "cp"
    ) )
AND ( LIMIT-TO ( SUBJAREA , "COMP"  ) )
AND ( LIMIT-TO ( LANGUAGE , "English" ) )
```

Fig. 5 Search query of the reference study.

the syntax of the respective platforms, were employed for the other databases. The search was executed on 16 July 2024, with no additional time constraints specified in the query formulation.

The selection process in the reference study was guided by the following inclusion criteria: (a) publication in a journal or conference proceedings, (b) related to computer science, and (c) published in English. The exclusion criteria comprised: (a) duplicated articles, (b) articles that presented results of surveys or reviews, (c) articles not related to using vision-based methods, and (d) articles not related to human AQA.

The results of the inclusion and exclusion steps of the PRISMA methodology are summarized in the first column of Table 5, directly extracted from [33]. At the end of the workflow, 96 studies were included in the systematic review. Among these, two were identified by the authors through a backward reference search of the selected papers. An internal verification subsequently revealed that these two papers corresponded to updated versions of studies already present in the set obtained from the three bibliographic databases. To ensure consistency and restrict comparisons to results derived from the three engines, these two additional articles were excluded from the reference set used as the benchmark for the evaluation. Consequently, the remaining 94 articles were considered the reference corpus for evaluation.

From these studies, 47 AQA datasets were identified and extracted. Datasets explicitly reported as *self-created* in the reference study were excluded from the comparison, as they are not publicly available and cannot be independently reproduced. Both the selected papers and the corresponding datasets constitute the basis for comparison with the results produced by the proposed framework.

4.2 Replication of the reference case study

To establish a baseline, we replicated the identification phase of the reference study, adding temporal boundaries to account for the period elapsed since the original publication. The remaining phases were not replicated, as they rely on subjective human judgment that cannot be systematically reproduced from the information reported in the original study, a constraint inherent to all manual systematic reviews. The replicated search returned 245 documents, reduced to 152 after duplicate removal. These results are detailed in the second column of Table 5. As discussed in Section 5.1, minor differences with respect to the original figures are consistent with the known temporal variability of bibliographic database indices and do not compromise the validity of the comparison.

4.3 Framework implementation and toolchain

The experimental evaluation employed LLMs from the Gemini 2.5 family [14] for both the screening and full-text eligibility phases. `gemini-2.5-flash` [22] was selected as the primary inference engine for large-scale tasks due to its favorable trade-off between speed, cost, and output accuracy. Its use enables efficient processing of large collections of papers while maintaining adequate accuracy. `gemini-2.5-pro` [23] was reserved for harmonization steps that require stronger reasoning capabilities and higher accuracy. The combination of these two LLMs supports a scalable architecture in which higher-capacity inference is applied only where it is most beneficial. Although this experimental validation employs a specific LLM, the proposed framework is inherently model-agnostic. The pipelines are directly compatible with other state-of-the-art models, such as GPT-4 [1] or the Claude 4 family [4], enabling the future integration of advanced paradigms like chain-of-thought reasoning.

To ensure fine-grained control over response determinism and minimize variability across repeated executions, all LLM invocations were performed under decoding configurations designed to approximate deterministic behavior. For the selected LLMs, this control is primarily achieved through the temperature parameter, which governs randomness in token selection by modulating the probability distribution over candidate tokens; as temperature decreases, probability mass becomes more concentrated on high-likelihood tokens, reducing output variability. In the limiting case, setting $T = 0$ yields greedy decoding, in which the highest-probability token is selected at each generation step. Determinism was further reinforced through Top-k and Top-p constraints: Top-k sampling restricts token selection to the k most probable candidates, while Top-p (nucleus) sampling restricts selection to the smallest candidate set whose cumulative probability mass exceeds a threshold p. In our experiments, decoding parameters were set to $T = 0$, Top-$k = 1$, and Top-$p = 1$, substantially reducing variability in the generated outputs.

It is important to note, however, that strict determinism cannot be fully guaranteed in LLM deployments, even under greedy decoding. Factors such as distributed execution environments and limited control over internal random seeds may introduce minor run-to-run differences. Achieving full determinism would generally require local deployment with complete control over random seeds and execution environments, typically at the cost of access to larger-capacity models.

For the full-text eligibility assessment, documents were encoded in dense vector embeddings using the Gemini text embedding model (specifically, `gemini-embedding-2`) [24]. This embedding model was selected for its ability to produce compact and semantically informative representations suitable for similarity-based retrieval. All embeddings were indexed in a Facebook AI Similarity Search (FAISS) vector store [16, 28], chosen due to its maturity, robustness, and ability to handle high-dimensional vector data with low latency, making it ideal for the workflow.

All experimental workflows were implemented on a personal computer equipped with 32 GB of RAM, a 1 TB SSD, and a 14-core Intel Core i7 CPU running at 3.1 GHz, with Windows 11 as the operating system. The codebase was developed in Python 3.11, chosen for its widespread adoption in the scientific community, extensive library ecosystem, and ease of customization. The entire workflow was

```
TITLE-ABS-KEY ("action quality" OR "action assessment" OR "
    action evaluation" OR AQA)
AND TITLE-ABS-KEY (dataset OR "data set" OR database OR "data
    catalogue" OR "data repository" OR "data sharing" OR "open
     data")
AND PUBYEAR > 2012 AND PUBYEAR < 2025
AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO ( DOCTYPE , "cp"
     ) )
AND ( LIMIT-TO ( SUBJAREA , "COMP" ) )
AND ( LIMIT-TO ( LANGUAGE , "English" ) )
```

Fig. 6 Search query of the proposed framework.

orchestrated using `LangChain` as a unified framework. `pandas` was used for structured data handling, while `pydantic` was used to define validated input and output schemas. PDF documents were loaded using `PyPDFLoader` and segmented into chunks with `RecursiveCharacterTextSplitter`. Interactions with Gemini models were handled through the `google.genai` client library. Finally, text embeddings were generated using `GoogleGenerativeAIEmbeddings` and indexed with the `faiss` library.

4.4 Workflow execution in the case-study setting

This subsection describes the concrete execution of the proposed methodology within the experimental setting, explicitly mapping each step to the corresponding PRISMA phase in the reference study.

4.4.1 Identification

The proposed workflow starts from a query executed on the same database engines considered in the reference study. The query adopted in the proposed workflow is reported in Figure 6. Compared with the original query shown in Figure 5, our query is more permissive in referring to the research domain and is therefore expected to retrieve a broader set of results. In particular, only minimal and widely used terms were included to cover the possible variations used to refer to AQA. This choice accounts for the fact that the terminology associated with the domain has progressively evolved and become standardized over time, whereas earlier works often adopted heterogeneous formulations.

In addition, the query explicitly requires the presence of terms related to data sources. This constraint reflects the objective of the case study, which targets the identification and analysis of AQA datasets: by anchoring the search to articles that mention the use, release, or discussion of data, the query filters out purely methodological contributions that would not yield relevant dataset information.

Particular attention was paid to the definition of the reference period. The temporal window was restricted to publications between 2013 and 2024 (inclusive). The lower bound was set for demonstration purposes, as no relevant articles were identified prior to 2013, and the first study included in the reference review also dates back to that year. The upper bound was introduced to limit the query to studies published

up to the final year covered by the reference study. However, since the reference study executed its queries on 16 July 2024, all publications released after that date were excluded from the proposed workflow to ensure a fair and consistent comparison. For IEEE Xplore and Web of Science, this filtering was performed by exploiting the publication date metadata returned by the respective database engines. In contrast, Scopus does not consistently provide a reliable publication date field in the query metadata. Consequently, for Scopus results, a manual verification step was performed solely for comparison purposes, inspecting the publication dates reported in conference proceedings or journals. Given potential indexing delays affecting Scopus, a conservative strategy was adopted, excluding all papers published in July 2024 to avoid introducing a temporal advantage over the reference study. All other optional query constraints, including document type, subject area, and language, were kept consistent with those adopted in the reference study. No other optional filtering criteria were introduced.

4.4.2 Screening

The automated screening and subsequent harmonization of the extracted data sources were executed sequentially. The screening stage was driven by the prompt reported in Figure 7. In the prompt, the screening decisions are encoded as *yes*, *no*, and *maybe* to simplify LLM output parsing, corresponding respectively to the *include*, *exclude*, and *uncertain* labels defined in Section 3.2. It is important to note that the inclusion and exclusion criteria are equivalent to those of the reference study. However, criteria referring to the identification phase, such as duplication verification, were removed from the prompt to avoid redundancy with steps already performed upstream. On the other hand, an additional inclusion criterion was introduced to retain only articles that explicitly mention the use or production of datasets, in line with the dataset-focused objective of this case study. The corpus of articles was processed in fixed-size batches of 15 records. Batching was chosen as a practical trade-off between accuracy and efficiency. For each batch, the pipeline instantiates the screening prompt by inserting the `<Title>` and `<Abstract>` fields of the corresponding articles, enforces a rate-limit check before issuing the request, invokes the LLM, and then maps the parsed response back to the original records through the provided identifiers.

To ensure controlled and robust API usage, the implementation enforces a simple rate limiter that allows at most 2 requests within a sliding window of 60 seconds. This constraint increases total processing time but does not affect the methodological validity of the results, and would be substantially reduced in production deployments with higher API quotas. LLM calls are only executed when the current call count falls below this threshold. LLM outputs must conform to a structured JSON format. The returned responses are then validated against the corresponding Pydantic schema. To improve robustness, an exponential backoff retry policy is applied on failures, with an initial delay of 10 seconds, a doubling strategy across retries, and a maximum number of attempts. If a batch cannot be successfully processed after all retries, the pipeline adopts a conservative fallback strategy: when responses cannot be parsed or are missing after all retry attempts, the affected records are assigned an indeterminate outcome (`decision = uncertain`).

Task. Evaluate each record based on the following Inclusion and Exclusion Criteria and return ONLY JSON that matches the provided schema.

Inclusion Criteria:

(a) The article mentions the use or production of datasets, databases, or other data sources.
(b) The article is related to computer science.

Exclusion Criteria:

(a) The article states that the data are not available, must be requested from the authors or are not free (cost is required).
(b) The article presents results of surveys or reviews.
(c) The article is not related to using vision-based methods.
(d) The article is not related to Human Action Quality Assessment (AQA).

Constraint. ALL the requirements should be met.

Output Instructions:

- For each provided record, return an object with: id, decision ('yes'—'no'—'maybe'), reason, data_source.
- If the title or abstract is missing or empty, use decision='no' and reason='Title or abstract is missing', data_source=''.

Records to evaluate: <Title> and <Abstract> of each article.

Fig. 7 Prompt template used to instruct the LLM during the automated screening phase.

Following screening, dataset name harmonization is executed as a separate phase using the prompt shown in Figure 8. Unlike screening, this step does not employ batching; however, the same rate-limiting, retry, and output validation policies described above are applied.

4.4.3 Eligibility

The eligibility phase uses the same operational policies as screening (rate limiting, LLM model and decoding parameters, retry/backoff, and Pydantic validation). However, unlike screening, articles are processed individually at the document level rather than in fixed-size batches, and a RAG system is introduced. The use of RAG at this stage is motivated by the fact that scientific articles often exceed the effective context window of the LLM, and relevant information such as dataset names may be scattered across distant sections. Retrieving the most relevant chunks before prompting allows the generative step to focus on the portions most likely to contain the information needed for eligibility decisions. This process is applied only to articles for which no dataset was extracted during the screening phase, to improve computational efficiency and reduce processing costs. Section 5.3.3 reports a dedicated sensitivity analysis that empirically quantifies the impact of this restriction on the completeness of the final dataset corpus.

Task. Check and harmonize all dataset names for consistency and correctness.

Guidelines:

- Remove descriptors like 'dataset', 'database', 'data'. Also remove unnecessary adjectives like 'open', 'new'.
- If multiple datasets are listed, leave the separation with ';' (semicolon) without spaces between different datasets.
- Always capitalize the first letter of each dataset. If the name is all uppercase, leave it unchanged.
- Prefer the recognized abbreviation if it exists, otherwise use the full name (*e.g.*, 'Massachusetts Institute of Technology' becomes 'MIT', 'Piano Skills Assessment' becomes 'PISA').
- Keep all actual dataset names; do NOT remove or discard any real dataset.
- If a field is empty, keep it empty.

Output Instructions:

- Return only JSON matching the schema.
- Return exactly one harmonized string per input row, in the same order as provided.

Records to harmonize: <Title> and <Abstract> of each article.

Fig. 8 Prompt template used to instruct the LLM to harmonize dataset names.

Full texts are downloaded directly from the respective sources and loaded as PDFs into a local repository. PDFs are then cleaned and split using a character-based splitter with a maximum chunk size of 2000 characters and an overlap of 400 characters, which helps preserve contextual continuity across adjacent chunks. Embeddings are computed with the configured embedding model and stored in a FAISS vector store, as explained previously.

For each document, the retrieval is driven by the fixed query "Human Action Quality Assessment (AQA) datasets", which is designed to focus on data sources relevant to AQA. A max-marginal-relevance (MMR) strategy is adopted rather than a standard similarity search [12, 29]. This approach is motivated by the need to balance relevance and diversity, reducing redundancy among chunks: MMR explicitly penalizes similar content and favors the selection of complementary information. The parameters that control the returned chunks are set proportionally to the total number of chunks, with $k = \lfloor N/3 \rfloor$ and $fetch\_k = \lfloor N/2 \rfloor$, where N denotes the total number of chunks in the index. Here, $fetch\_k$ defines the size of the initial candidate pool retrieved by semantic similarity search, while k specifies the number of chunks finally selected after MMR re-ranking. The trade-off between relevance and diversity is controlled by the MMR trade-off parameter λ. A high value ($\lambda = 0.9$) was adopted to strongly prioritize semantic relevance to the query while still introducing a small diversification effect.

Task. You are an assistant that analyzes full scientific papers.

Guidelines: Given the following extracted text from a paper, extract ONLY the official and recognized names of datasets that satisfy the following conditions:

(a) The dataset is specifically used for Human Action Quality Assessment (AQA). This condition must be strictly respected.
(b) Include only real, unique, and identifiable resources (dataset, database, or data collection).
(c) Do NOT write generic descriptions (*e.g.*, 'a large dataset', 'clinical dataset'); if the paper only describes the data generically, leave this field empty.

Output Instructions:

- Return valid JSON only, with a single field `"datasets"` that is a list of dataset names, *e.g.*, {`"datasets": ["FINA09","MTL-AQA"]`}.
- If multiple datasets are mentioned, include all of them in the same list.
- If no datasets meet these conditions, return an empty list.

Fig. 9 Prompt template used to instruct the LLM to extract dataset names from the full text of scientific papers.

For each paper, up to five candidate chunks are concatenated and passed to the LLM using the extraction prompt shown in Figure 9. Extracted dataset names are then normalized with the same harmonization prompt shown in Figure 8.

4.4.4 Inclusion

Only articles for which at least one dataset was identified were considered eligible and included in the systematic review, forming the final corpus for subsequent analysis. Finally, Figure 10 shows the number of articles identified for each stage of the workflow. Specifically, a total of 91 articles were included in the systematic review. From these studies, 65 unique AQA datasets were extracted and are compared in the following section.

5 Results and discussion

The evaluation examines the framework's ability to reproduce the reference study by retrieving relevant papers and associated datasets with coverage and relevance comparable to a manual review. The analysis focuses primarily on study selection and dataset identification, comparing the outputs of the automated pipeline with those reported in the reference study. It also considers the practical viability, robustness, and portability of the framework. In all cases, exact matches are not required; rather, the goal is to achieve similar coverage and relevance to that of the manual review.

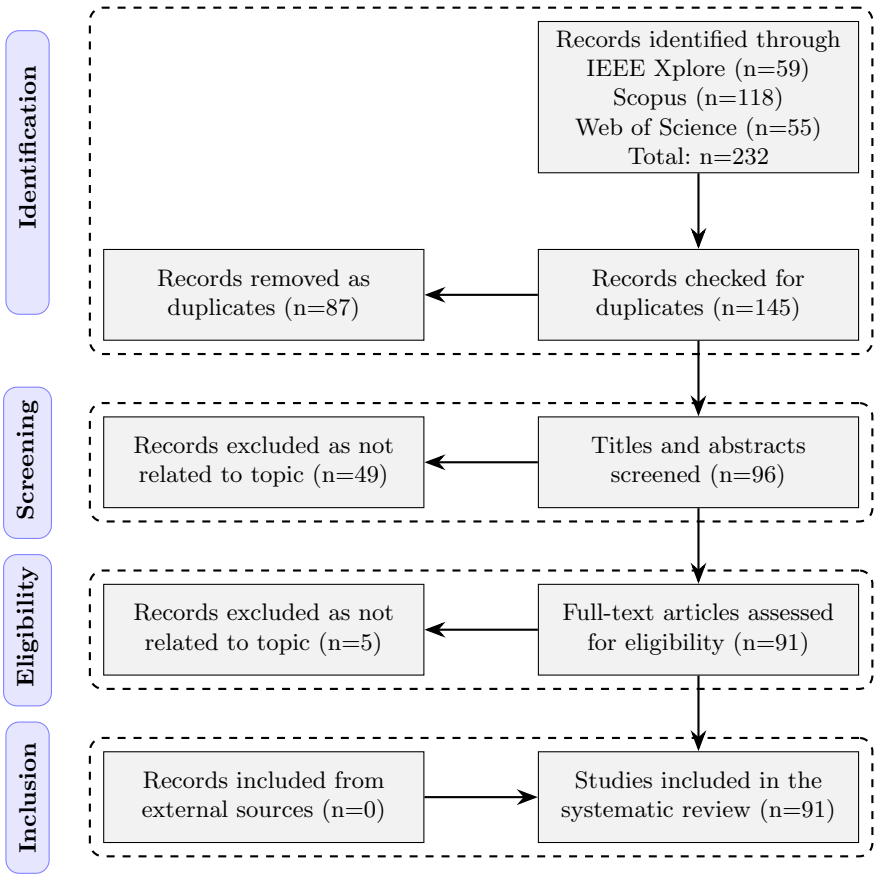

Fig. 10 Diagram of the PRISMA workflow for the proposed framework, where n denotes the number of papers.

5.1 Results on study identification

The proposed framework ultimately identifies 91 articles for inclusion in the systematic review, 67 of which are also present among the 94 articles in the reference study. This corresponds to a detection rate of 71.28%, computed after excluding the two studies added through backward reference search in the reference set, as previously discussed. Before analyzing the sources of this gap, it is important to distinguish two structurally different sources of missed records. The first is *query-design-induced*: articles that do not mention datasets in their title, abstract, or keywords are excluded by construction at the identification phase, independently of any LLM behavior. The second is *pipeline-induced*: articles that entered the automated stages but were incorrectly excluded by the screening or eligibility components. As shown in Table 5, the screening and eligibility stages contribute only marginally to the loss: the screening stage correctly excludes 49 irrelevant articles while retaining all reference-study records that reach it. Of the 145 records that entered the screening stage, 3 (approximately 2.1%) were labeled *uncertain* due to schema validation failures or low-confidence outputs,

and were subsequently resolved by manual review, consistent with the conservative fallback policy described in Section 3.2. The eligibility stage, in turn, removes only three reference articles, none of which contributes unique datasets. Of the 96 records that passed screening, a dataset had already been extracted directly from the title and abstract for 75 of them; therefore, only the remaining 21 records were processed by the RAG-based full-text eligibility pipeline. Overall, the automated stages retained 67 of the 70 reference studies that entered the pipeline, corresponding to a conditional retention rate of 95.71%. Therefore, of the 27 missing reference articles, 24 fall into the *query-design-induced* category and were excluded before any automated decision-making by the LLM and RAG components was applied, while only 3 are attributable to *pipeline-induced* decisions in the eligibility stage.

This finding suggests that the LLM-based components of the framework operate with high fidelity, while the main limitation lies upstream in the bibliographic search stage. The missing records can be explained by factors intrinsic to the identification phase. First, and most importantly, the query was intentionally designed to retrieve only articles that explicitly refer to datasets, since the primary objective of the review is dataset identification rather than broad methodological coverage. As a consequence, articles discussing AQA methods that do not mention data sources in the title, abstract, or keywords are excluded by design. Second, terminological heterogeneity in the AQA literature, especially in earlier works, where non-standardized formulations were more common, may have led some records to fall outside the query's lexical coverage. Third, the dynamic nature of bibliographic databases, including delayed indexing, metadata corrections, and changes in record availability over time, introduces a residual source of variability that cannot be fully controlled.

While the stringent query design limits the overall detection rate compared to a broad-scope manual search, the automated pipeline provides a crucial advantage: it replaces subjective human evaluation with a highly consistent and reproducible procedure. By standardizing the screening and eligibility stages, the framework reduces the variability between reviewers that typically affects manual workflows, ensuring that all inclusion and exclusion decisions are traceable and systematically aligned with the PRISMA guidelines. To confirm that this consistency is not an artifact of a single execution, we repeated the automated stages across five independent runs; the pipeline exhibited high inter-run agreement, with residual disagreements confined to genuine boundary cases, as detailed in Section 5.3.2.

A comparison with our replication of the reference study further highlights the inherent variance of bibliographic search results. Despite adopting the same declared queries, the replication produced different record counts at the identification stage. In particular, after duplicate removal, it yielded 27 more articles than the reference study. At the same time, neither corpus is a strict superset of the other: 92 of the 94 articles included in the final corpus of the reference study were also retrieved in the replicated study, leaving a discrepancy of two articles. We report these differences strictly to document the baseline variations introduced by search engine indexing and temporal factors before any automated screening occurs.

For completeness, the full list of the 91 articles identified by the proposed framework for inclusion in the systematic review is provided in A.1. This table details the

Table 5 Comparison of the PRISMA workflow between the reference study, the replicated study, and the proposed framework.

PRISMA Step	Reference Study	Replicated Study	Proposed Framework
Identification			
Records initially identified	230	245	232
IEEE Xplore	50	60	59
Scopus	119	131	118
Web of Science	61	54	55
Records removed as duplicates	105	93	87
Records after identification	125	152 (92)	145 (70)
Screening			
Records excluded during screening	24		49
Records after screening	101		96 (70)
Eligibility			
Records excluded during eligibility	7		5
Records after eligibility	94		91 (67)
Inclusion			
Records integrated from reference search	2		0
Final records included	96		91 (67)

Note: The replicated study reports only the Identification phase; empty cells denote steps not performed. Numbers in parentheses indicate the records in common with the final set of articles included in the reference study, excluding the two studies identified by the authors through backward reference search.

bibliographic references (DOI, year, title, and authors) along with the corresponding datasets extracted by the automated pipeline for each article.

5.2 Results on dataset identification

The proposed framework identified 40 of the 47 AQA datasets reported in the reference study, corresponding to an extraction recall of 85.11%. A quantitative and qualitative synthesis of the discrepancies between the two studies is provided in Table 6, which compares the 7 datasets missed by the framework with the 25 newly identified ones. To understand the nature of these differences while reducing the risk of confirmation bias, a multi-annotator validation was conducted on all 32 discrepant datasets across both groups. Specifically, each resource was analyzed based on its relationship to the AQA domain, its availability status (public vs. private), and the inclusion status of related articles. Furthermore, the scientific impact was assessed by collecting citation counts from Google Scholar (as of December 20, 2025). This metric was chosen to evaluate whether the automated framework can retrieve influential resources that might have been overlooked in the original review, or vice versa. Crucially, this citation analysis was performed on two levels: we examined both the article identified during the workflow and the original publication where the dataset was first introduced. This dual perspective enables a more robust evaluation of the framework's ability to identify relevant sources.

Table 6 Summary comparison between datasets missed by the proposed framework and newly identified datasets.

Category	Feature	Missed by Framework (A.2)	Newly Identified (A.3)
Quantity	Total Datasets	7	25
Relation	AQA-ready	6	11
	AQA-related	1	14
Availability	Public	3	14
	Private	4	11
Inclusion	Found paper, missed dataset [a]	1 (14.3%)	8 (32.0%)
	Paper not found [b]	6 (85.7%)	17 (68.0%)
Impact	Total Citations (Articles)	197	967
	Avg. Citations (Articles)	**28.14**	**38.68**
	Total Citations (Original)	64	27,413
	Avg. Citations (Original)	**21.33**	**1,713.31**

Note: **a** The article is present in the study's final collection, but the specific dataset was not extracted. For newly identified datasets, this means that the article was included in the reference study, but the dataset name was omitted. **b** The article was not retrieved during the identification phase of the respective study/framework. The replicated version is used as the reference study.

As part of this manual validation, all 65 extracted dataset names were inspected against their corresponding source publications. No instances of fabricated or non-existent dataset names were identified: every extracted name corresponds to a real, verifiable resource. The main source of residual error was not hallucination, but naming variance, since the same dataset is sometimes referred to with different conventions across different papers, such as different abbreviations or formatting. The harmonization step described in Section 3.2 substantially reduced this variance, though a small number of cases required additional manual reconciliation to a common nomenclature, since automated harmonization did not fully converge on a single canonical form for every variant.

A primary focus of this validation was the relation of each dataset to the domain. The resources have been classified into three levels of suitability:

- AQA-ready: Datasets fully prepared for AQA, containing ground truth quality scores, error labels, or clinical metrics required for action assessment;
- AQA-related: Datasets not natively AQA-ready but adaptable (*e.g.*, action recognition or motion analysis) with additional annotations;
- Not Applicable to AQA: Resources lacking the necessary quality variance or temporal granularity required for adaptation to AQA, belonging to an unaligned domain, or those not belonging to the dataset category (*e.g.*, methods or tools).

Three annotators independently assigned each of the 32 discrepant datasets to one of these three categories. Inter-rater reliability was quantified using Fleiss' κ [17], yielding $\kappa = 0.7917$ for the AQA suitability classification, which indicates substantial agreement. Disagreements were reviewed through discussion, and final labels were assigned by majority vote among the three annotators. The consensus validation confirmed the high precision of both workflows, as no *"Not Applicable to AQA"* datasets were included in either the reference study or the proposed framework. Notably, while almost all datasets from the reference study not captured by our framework were classified as *AQA-ready* (6 out of 7), the proposed framework introduced a more balanced distribution, identifying both *AQA-ready* (11) and *AQA-related* (14) resources.

Regarding dataset availability, the two systems performed similarly, extracting a comparable mix of public and private datasets. As shown in the summary table, most undetected datasets (85.7%) are associated with articles that the pipeline failed to retrieve during the initial identification phase. Conversely, the automated pipeline identified a substantial number of additional resources not included in the original review. Notably, 32% of these newly identified datasets were extracted from articles already included in the reference study, but whose dataset names had been omitted from the mapping. Moreover, these newly identified datasets exhibit higher citation impact, particularly compared with the original papers. For a detailed breakdown of these findings, including individual metadata and citation counts, the reader is referred to A: A.1 provides the full list of papers included in the systematic review, A.2 lists the datasets missed by the framework, while A.3 summarizes those newly identified.

5.3 Robustness and portability analyses

Beyond the core identification and dataset-extraction results, we assess the practical viability and generalizability of the framework along four complementary dimensions: computational cost, run-to-run output stability, sensitivity to the RAG-based extraction restriction, and portability across LLM families.

5.3.1 Computational cost and processing time

Beyond retrieval accuracy, the practical viability of an automated systematic review framework depends heavily on its computational and economic footprint. To systematically evaluate both aspects, we conducted five independent, supplementary executions ($n = 5$) of the screening and eligibility phases under identical configurations ($T = 0$, Top-$k = 1$, Top-$p = 1$). It is important to note that these runs are in addition to the primary execution whose findings are reported in the preceding sections of this paper. For the eligibility phase, each of the five runs was executed on the fixed set of records produced by the primary screening run, rather than on the output of the corresponding screening replicate; this isolates the variance attributable to eligibility-stage processing from any variance originating upstream in screening.

Table 7 reports the aggregated resource consumption and estimated API costs for each stage of the experimental validation. The complete pipeline required a total of 549.9 seconds (approximately 9.2 minutes) of execution time and incurred an estimated cost of $0.13. A significant portion of this cost ($0.0605, approximately 47%) is attributable to the vectorization of full-text documents for the FAISS index. The

Table 7 Computational performance and estimated costs across pipeline stages.

Stage	Model	Calls	In-Tokens	Out-Tokens	Time (s)	Cost (USD)
Screening (145 records)						
Screening	Flash 2.5	11	$42,634$	$9,820 \pm 343$	305.6 ± 8.1	0.0373 ± 0.0009
Harmonization	Pro 2.5	1	$2,601 \pm 8$	$1,016 \pm 10$	56.4 ± 13.3	0.0134 ± 0.0001
Subtotal	–	**12**	**45,235**	**10,836**	**362.0**	**0.0507**
Eligibility (21 records)						
Vector Indexing	Embedding 2	1	$302,586$	–	49.9 ± 2.3	0.0605
Extraction (RAG)	Flash 2.5	20	$46,083$	225 ± 6	114.2 ± 5.5	0.0144
Harmonization	Pro 2.5	1	511 ± 5	191 ± 5	23.8 ± 6.4	0.0025 ± 0.0001
Subtotal	–	**22**	**349,180**	**416**	**187.9**	**0.0774**
Total	–	**34**	**394,415**	**11,252**	**549.9**	**0.1282**

Note: Values are reported as Mean ± Standard Deviation. Metrics with zero variance are reported as exact values.

screening and extraction phases account for only $0.0517 in total, confirming that the cost per record for the main LLM-based components is minimal.

The architecture proved effective in balancing cost and performance. Consistent with the design described in Section 4.3, `gemini-2.5-flash` accounted for the majority of token consumption (over 88% of total input tokens), while `gemini-2.5-pro` was invoked selectively for harmonization. Notably, despite processing full-text documents via the RAG mechanism described in Section 4.4.3, the eligibility stage required only $0.0144 for the processed papers, less than half the cost of the main screening stage. This low cost per document is a direct consequence of the RAG retrieval mechanism, which limits the LLM context to a small set of semantically relevant chunks rather than the entire document. It is also worth noting that, among the 21 records entering the eligibility phase, only 20 API calls were issued: one article received no chunks from the MMR retrieval step, as no segments from that document were ranked within the selected retrieval window.

5.3.2 Output variability across repeated executions

A key concern regarding the integration of LLMs into systematic review pipelines is the potential for non-deterministic outputs to undermine reproducibility. To empirically characterize this risk, we analyzed the output variance across the five supplementary runs introduced in Section 5.3.1.

Screening phase. The inclusion/exclusion decisions were analyzed across all five runs by computing Fleiss' κ, treating each execution as a separate rater. Of the 145 screened records, 142 received a unanimous decision across all runs, yielding a Fleiss' $\kappa = 0.9726$, indicating almost perfect agreement. The three discrepant records represent genuine boundary cases whose abstracts discuss action recognition methods without unambiguous AQA relevance, causing the model to oscillate between inclusion

and exclusion across runs; if multiple runs are possible, such cases would be appropriately marked as *uncertain* and routed to manual review by the conservative fallback policy described in Section 3.2.

Dataset extraction from abstracts showed comparable stability: 141 of 145 records (97.24%) produced identical outputs across all runs. The four discrepancies fall into two categories: in one case, a dataset was extracted in only one run; the remaining three correspond to the same boundary-case records that triggered inclusion discrepancies, where the extraction output was consistently aligned with the upstream screening decision (*i.e.*, datasets were successfully extracted whenever the record was classified as *included*). Crucially, a further evaluation confirmed that the dataset missed during abstract screening is consistently recovered during eligibility.

Eligibility phase. To isolate the variance of the RAG-based extraction stage independently of upstream screening variability, the five eligibility runs used the fixed output of the primary screening run as input. Of the 21 processed full-text records, 19 (90.5%) produced identical harmonized dataset lists across all runs. Both discrepant records involved peripheral secondary datasets rather than primary evaluation datasets, which remained consistent across all runs.

Taken together, these results confirm that the zero-temperature decoding configuration, combined with schema-constrained structured outputs and automated harmonization, effectively constrains LLM variability to boundary cases that would appropriately require human review regardless of the level of automation adopted. We note, however, that five runs constitute an initial empirical estimate rather than an exhaustive characterization of the variance distribution; a more comprehensive stability analysis across a larger number of runs and alternative model configurations remains a valuable direction for future work. The sources of discrepancy are detailed in B.1.

5.3.3 Sensitivity to RAG extraction constraints

To empirically assess the impact of restricting full-text RAG-based extraction to articles for which no dataset was extracted during screening, we applied the RAG-based eligibility procedure to the complementary subset, i.e., all 75 articles exempted from this step because a dataset had already been identified from title and abstract.

Full-text analysis recovered additional dataset mentions in 10 of the 75 articles (13.3%), confirming that full-text analysis can surface information not captured by title/abstract screening alone; a per-article breakdown is provided in B.2. Across these 10 articles, 12 additional dataset mentions were recovered in total. Cross-referencing them against the original corpus of unique datasets (Section 5.2) showed that 11 corresponded to datasets already present elsewhere in the corpus, extracted from at least one other included article. The remaining mention (*Waseda Backsquat*), recovered from a single article's full text, did not appear elsewhere in the corpus. Manual verification by all annotators confirmed that this dataset is indeed a valid, publicly documented *AQA-ready* dataset, ruling out the possibility of a hallucinated extraction.

Extending full-text RAG to the entire screening-positive subset would therefore have increased the final corpus from 65 to 66 unique datasets (+1.5%). We consider this a small but non-negligible impact: it confirms that the restriction is a reasonable

trade-off between computational efficiency and completeness for the purposes of this case study, while also showing that the impact is not zero. This motivates our recommendation that full-text RAG-based extraction be applied to the entire screened corpus when the goal is to maximize dataset recall.

5.3.4 Cross-model portability analysis

To assess whether the results of the proposed framework depended on the LLM family adopted in the primary evaluation, we performed an additional end-to-end execution using Claude. Mirroring the two-tier model allocation adopted for the primary Gemini-based configuration (Section 4.3), `claude-sonnet-4-5` [3] was used for the main screening and eligibility pipeline, while `claude-opus-4-5` [2] was reserved for the harmonization step requiring stronger reasoning capabilities. The embedding model used to index full-text articles for the RAG-based eligibility phase, `gemini-embedding-2` [24], was kept unchanged across both executions to isolate the effect of the generative model from that of the retrieval component. Both executions started from the same set of 145 records obtained after identification and deduplication. To ensure comparability, the corpus, eligibility criteria, prompt templates, structured output schemas, fallback policies, and RAG configuration were kept unchanged, while only the model-specific API integration and supported inference parameters were adapted. Since this supplementary experiment consisted of a single execution, it was designed as a model-portability assessment rather than as an additional analysis of run-to-run stability.

Table 8 compares the study-selection outcomes obtained with Gemini and Claude. At the screening stage, Gemini retained 96 records and Claude retained 95, with 94 records retained by both models. After the eligibility phase, the final corpora contained 91 and 87 studies, respectively, of which 84 were shared. Gemini recovered 67 of the 94 reference studies, corresponding to a detection rate of 71.28%, whereas Claude recovered 64, corresponding to 68.09%. Of these, 61 reference studies were recovered by both executions, while six were identified only by Gemini and three only by Claude. These results indicate a high degree of agreement between the two executions, although Claude adopted a slightly more restrictive behavior across the selection stages.

Table 9 reports the corresponding dataset-extraction results. Gemini identified 65 unique datasets, compared with 57 identified by Claude, with an overlap of 54 datasets. Of the 47 datasets reported in the reference study, Gemini recovered 40 and Claude recovered 38, corresponding to extraction recalls of 85.11% and 80.85%, respectively. The 38 reference datasets identified by Claude were also recovered by Gemini. In addition, Gemini identified 25 datasets not reported in the reference study, while Claude identified 19; 16 of these additional resources were shared by both executions. All additional datasets were classified as *AQA-ready* or *AQA-related* based on the rubric defined in Section 2. Overall, the substantial overlap at both the study and dataset levels provides preliminary evidence that the framework can be transferred across different LLM families without modifying its methodological components. However, the slightly lower coverage obtained with Claude and the use of a single supplementary execution indicate that these findings should be interpreted as evidence of operational portability rather than definitive model-independent robustness.

Table 8 Comparison of the PRISMA workflow between the Gemini- and Claude-based executions.

PRISMA Step	Gemini	Claude	Gemini ∩ Claude
Screening			
Records entering screening	145	145	145
Records excluded during screening	49	50	–
Records after screening	96	95	94
Eligibility			
Records entering the RAG pipeline	21	29	20
Records excluded during eligibility	5	8	–
Records after eligibility	91	87	84
Inclusion			
Final records included	91	87	84
Reference-study records recovered	67	64	61
Reference-study detection rate	71.28%	68.09%	64.89%

Note: The "Gemini ∩ Claude" column reports records shared by both executions.

Table 9 Comparison of dataset extraction results between the Gemini- and Claude-based executions.

Metric	Gemini	Claude	Gemini ∩ Claude
Unique datasets extracted	65	57	54
Reference datasets recovered	40	38	38
Reference-dataset extraction recall	85.11%	80.85%	80.85%
Additional datasets identified	25	19	16

Note: The "Gemini ∩ Claude" column reports datasets identified by both executions.

The supplementary Claude-based execution required 448.7 seconds (approximately 7.5 minutes) and incurred an estimated cost of $0.6759, including $0.0780 for embedding generation and vector indexing. Compared with the Gemini-based configuration, Claude completed the workflow faster (448.7 vs. 549.9 seconds) but at a substantially higher total cost ($0.6759 vs. $0.1282). Since the Claude results derive from a single execution, this comparison is descriptive and should not be interpreted as an estimate of run-to-run variability.

For a detailed breakdown of these findings, the reader is referred to B: B.3 lists the study-level disagreements between the two executions, B.4 summarizes the corresponding dataset-level disagreements, and B.5 reports the stage-level API calls, token consumption, execution time, and estimated costs.

6 Conclusions

This work demonstrates that LLMs and RAG can be integrated into a PRISMA-aligned workflow to automate systematic review tasks without sacrificing transparency or reproducibility. We establish this claim through a single-domain case study rather than a multi-domain evaluation, so the present evidence demonstrates feasibility and

internal reproducibility rather than cross-domain generality. The experimental validation showed that the primary bottleneck lies not in the automated components, but in the bibliographic identification phase, a finding that has direct implications for the design of future automated review pipelines. This is evident in the retrieval results: the framework retrieved 91 articles, including 67 of the 94 studies reported in the reference review (71.28% detection rate). By explicitly disentangling the sources of missed articles, we observe that the screening and eligibility components introduced only a *pipeline-induced* marginal loss (retaining 95.71% of the records that reached them). The primary bottleneck was therefore the *query-design-induced* constraint rather than the LLM stages. At the dataset level, it extracted 65 distinct datasets, including 40 of the 47 reference datasets (85.11% extraction recall) and 25 additional datasets not reported in the original review. The newly discovered datasets further highlight the framework's potential to surface relevant information not captured in the reference review. Repeated executions confirmed that the pipeline behaves near-deterministically, with 142 of 145 screening decisions unanimous across five runs (Fleiss' $\kappa = 0.9726$); the few divergent cases were genuine boundary records that the conservative fallback policy routes to manual review.

This study has a number of limitations that should be noted. The framework was validated on a single domain (AQA), and it remains unclear how well it would transfer to fields with substantially different terminological conventions, publication structures, or data annotation practices. Moreover, the observed detection rate means that over a quarter of the reference articles were missed, largely as a consequence of the dataset-oriented query design adopted in the case study. Broader query formulations could improve recall, though likely at the expense of precision. Similarly, restricting full-text RAG-based extraction to articles without a dataset identified during screening cost one unique dataset. This motivates our recommendation that it be applied to the entire screened corpus when the goal is to maximize dataset recall. On the modeling side, the primary evaluation and stability analysis relied on the Gemini 2.5 family under one configuration, although the supplementary Claude execution provided preliminary evidence of cross-model portability. Lastly, although the computational footprint of the pipeline was quantified in terms of execution time and API cost, no quantitative evidence on human time savings or workload reduction was collected, as this would require a controlled comparison against manual review effort.

The transition toward semi-automated systematic reviews introduces critical ethical and governance challenges. Delegating eligibility and extraction tasks to LLMs raises valid concerns regarding algorithmic accountability, particularly concerning false negatives (unjustified exclusions) and the potential amplification of embedded model biases. To mitigate these risks, the framework enforces traceability by requiring the LLM to generate explicit, auditable justifications for every single inclusion and exclusion decision. By continuously logging these reasons alongside system configurations, the framework's reporting structure aligns with the disclosure principles advocated by emerging AI governance frameworks such as PRISMA-trAIce [26] and L-PRISMA [45], which respectively emphasize transparent reporting of AI-specific design choices and explicit documentation of model provenance and verification procedures. Ultimate accountability for inclusion and exclusion decisions remains with

the human researchers, who can audit and override any automated decision through the logged justifications.

These results show that the framework can support systematic reviews without compromising methodological rigor while providing a complementary tool for evidence synthesis and targeted information retrieval. Beyond this case-study validation, the proposed pipeline is intended to be adaptable across domains by re-specifying the query template, eligibility criteria, and extraction schema while preserving the same PRISMA-aligned process. Future work will target three main directions. First, the bibliographic identification phase will be refined through query expansion and the integration of additional bibliographic engines to improve recall. Second, the RAG-based retrieval and metadata harmonization components will be optimized, and the single-execution portability assessment reported here will be extended into a systematic benchmark of additional LLM families across detection performance, extraction quality, and computational efficiency. Third, the framework will be validated in diverse research domains, which is essential to substantiate the transferability of the proposed approach. By making systematic review procedures more accessible, reproducible, and scalable, the proposed framework aims to lower the barrier for evidence synthesis in research domains where manual reviews remain prohibitively resource-intensive.

Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The datasets analysed during the current study are available in the Zenodo repository: https://doi.org/10.5281/zenodo.20137559.

Materials availability

Not applicable.

Code availability

The code used during the current study is available from the corresponding author on reasonable request.

Author contributions

Omitted for double-blind review. Full author contributions are provided in the separate Title Page.

References

[1] Achiam J, Adler S, Agarwal S, et al (2023) Gpt-4 technical report. arXiv preprint arXiv:230308774

[2] Anthropic (2025) Claude opus 4.5 system card. https://www.anthropic.com/claude-opus-4-5-system-card, accessed: 2026-07-14

[3] Anthropic (2025) Claude sonnet 4.5 system card. https://www.anthropic.com/claude-sonnet-4-5-system-card, accessed: 2026-07-14

[4] Anthropic (2025) Introducing claude 4. https://www.anthropic.com/news/claude-4, accessed: 2026-07-14

[5] Bastian H, Glasziou P, Chalmers I (2010) Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS medicine 7(9):e1000326

[6] Bolanos F, Salatino A, Osborne F, et al (2024) Artificial intelligence for literature reviews: opportunities and challenges: F. bolaños et al. Artificial Intelligence Review 57(10):259

[7] Borah R, Brown AW, Capers PL, et al (2017) Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. BMJ open 7(2):e012545

[8] Brîncoveanu C, Carl KV, Witzki A, et al (2026) Augmenting systematic literature reviews: A human-ai collaborative framework. In: German Conference on Artificial Intelligence (Künstliche Intelligenz), Springer, pp 3–17

[9] Brown A, Roman M, Devereux B (2025) A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. arXiv preprint arXiv:250806401

[10] Cao C, Arora R, et al (2025) Automation of systematic reviews with large language models. medRxiv pp 2025–06

[11] Cao C, Sang J, Arora R, et al (2025) Development of prompt templates for large language model–driven screening in systematic reviews. Annals of Internal Medicine 178(3):389–401

[12] Carbonell J, Goldstein J (1998) The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in

information retrieval, pp 335–336

[13] Clark J, Barton B, Albarqouni L, et al (2025) Generative artificial intelligence use in evidence synthesis: A systematic review. Research Synthesis Methods pp 1–19

[14] Comanici G, Bieber E, Schaekermann M, et al (2025) Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:250706261

[15] Delgado-Chaves FM, Jennings MJ, Atalaia A, et al (2025) Transforming literature screening: The emerging role of large language models in systematic reviews. Proceedings of the National Academy of Sciences 122(2):e2411962122

[16] Douze M, Guzhva A, Deng C, et al (2025) The Faiss library. IEEE Transactions on Big Data

[17] Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological bulletin 76(5):378

[18] Flemyng E, Noel-Storr A, Macura B, et al (2025) Position statement on artificial intelligence (ai) use in evidence synthesis across cochrane, the campbell collaboration, jbi and the collaboration for environmental evidence 2025. Cochrane Database of Systematic Reviews (10)

[19] Forero DA, Abreu SE, Tovar BE, et al (2025) Large language models and the analyses of adherence to reporting guidelines in systematic reviews and overviews of reviews (PRISMA 2020 and PRIOR). Journal of Medical Systems 49(1):80

[20] Galli C, Gavrilova AV, Calciolari E (2025) Large language models in systematic review screening: opportunities, challenges, and methodological considerations. Information 16(5):378

[21] Gao Y, Xiong Y, Gao X, et al (2023) Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:231210997 2(1)

[22] Google (2025) Gemini 2.5 flash. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash, accessed: 2026-07-14

[23] Google (2025) Gemini 2.5 pro. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro, accessed: 2026-07-14

[24] Google (2026) Gemini text embeddings api reference. URL https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/embedding-2, accessed: 2026-07-14

[25] Harasgama S, Pearce H, Appel C, et al (2026) Artificial intelligence tools for automating evidence synthesis: Scoping review. Journal of Medical Internet Research 28:e81597

[26] Holst D, Moenck K, Koch J, et al (2025) Transparent reporting of ai in systematic literature reviews: Development of the prisma-traice checklist. JMIR AI 4:e80247

[27] Huotala A, Kuutila M, Ralph P, et al (2024) The promise and challenges of using llms to accelerate the screening process of systematic reviews. In: Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering, pp 262–271

[28] Johnson J, Douze M, Jégou H (2019) Billion-scale similarity search with GPUs. IEEE Transactions on Big Data 7(3):535–547

[29] Karpukhin V, Oguz B, Min S, et al (2020) Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp 6769–6781

[30] Khalil H, Ameen D, Zarnegar A (2022) Tools to support the automation of systematic reviews: a scoping review. Journal of Clinical Epidemiology 144:22–42

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

[32] Lewis P, Perez E, Piktus A, et al (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems 33:9459–9474

[33] Liu J, Wang H, Stawarz K, et al (2025) Vision-based human action quality assessment: A systematic review. Expert Systems with Applications 263:125642

[34] Malik FS, Terzidis O (2025) A hybrid framework for creating artificial intelligence-augmented systematic literature reviews. Management Review Quarterly pp 1–27

[35] Marshall IJ, Kuiper J, Wallace BC (2016) RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. Journal of the American Medical Informatics Association 23(1):193–201

[36] Miao Y, Zhao Y, Luo Y, et al (2025) Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: Scoping review. Journal of Medical Internet Research 27:e80557

[37] Moher D, Liberati A, Tetzlaff J, et al (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Bmj 339

[38] Oami T, Okada Y, Nakada Ta (2024) Performance of a large language model in screening citations. JAMA network open 7(7):e2420496–e2420496

[39] Ouzzani M, Hammady H, Fedorowicz Z, et al (2016) Rayyan—a web and mobile app for systematic reviews. Systematic reviews 5(1):210

[40] Page MJ, McKenzie, et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372

[41] Pérez J, Díaz J, Garcia-Martin J, et al (2020) Systematic literature reviews in software engineering—enhancement of the study selection process using cohen's kappa statistic. Journal of Systems and Software 168:110657

[42] Rethlefsen ML, Kirtley S, Waffenschmidt S, et al (2021) PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Systematic reviews 10(1):39

[43] Scherbakov D, Hubig N, Jansari V, et al (2025) The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. Journal of the American Medical Informatics Association 32(6):1071–1086

[44] Scotti KL, Young S, Gainey MA, et al (2025) Artificial intelligence and automation in evidence synthesis: An investigation of methods employed in cochrane, campbell collaboration, and environmental evidence reviews. Cochrane Evidence Synthesis and Methods 3(5):e70046

[45] Shailendra S, Kadel R, Sharma A, et al (2026) L-prisma: An extension of prisma in the era of generative artificial intelligence (genai). Authorea Preprints

[46] Susnjak T (2023) Prisma-dfllm: An extension of prisma for systematic literature reviews using domain-specific finetuned large language models. arXiv preprint arXiv:230614905

[47] Thode L, Iftikhar U, Mendez D (2025) Exploring the use of llms for the selection phase in systematic literature studies. Information and Software Technology p 107757

[48] de la Torre-López J, Ramírez A, Romero JR (2023) Artificial intelligence to automate the systematic review of scientific literature. Computing 105(10):2171–2194

[49] Van De Schoot R, Bruin D, et al (2021) An open source machine learning framework for efficient and transparent systematic reviews. Nature machine intelligence 3(2):125–133

[50] Wagner G, Lukyanenko R, Paré G (2022) Artificial intelligence and the conduct of literature reviews. Journal of Information Technology 37(2):209–226

[51] Wang Z, Nayfeh T, Tetzlaff J, et al (2020) Error rates of human reviewers during abstract screening in systematic reviews. PloS one 15(1):e0227742

[52] Wang Z, Cao L, Danek B, et al (2025) Accelerating clinical evidence synthesis with large language models. npj Digital Medicine 8(1):509

[53] Zhao WX, Zhou, et al (2023) A survey of large language models. arXiv preprint arXiv:230318223 1(2)

Appendix A

This appendix provides the complete study-level and dataset-level results of the review. It reports all included papers and extracted data sources, together with the datasets identified exclusively by either the reference study or the proposed framework.

Table A.1: Overview of all identified papers with all datasets extracted.

DOI	Year	Title	Authors	Data Source
10.1007/978-3-319-10599-4_36	2014	Assessing the quality of actions	Pirsiavash, H.; Vondrick, C.; Torralba, A.	MIT-Skate; MIT-Dive
10.1109/CVPRW.2017.16	2017	Learning to Score Olympic Events	P. Parmar; B. T. Morris	FINA09
10.1109/SSCI.2017.8285270	2017	The open online repository of karate motion capture data: A tool for scientists and sport educators	Hachaj, T.; Ogiela, M.R.; Piekarczyk, M.	self-created
10.1007/978-3-030-00767-6_12	2018	End-to-end learning for action quality assessment	Li, Y.; Chai, X.; Chen, X.	MIT-Skate; UNLV-Dive; UNLV-Vault
10.1016/j.patcog.2017.12.007	2018	Motion analysis: Action detection, recognition and evaluation based on motion capture data	Patrona, F.; Chatzitofis, A.; Zarpalas, D.; Daras, P.	MSR-Action3D; MSRC-12
10.1109/ICIP.2018.8451364	2018	S3D: Stacking Segmental P3D for Action Quality Assessment	X. Xiang; Y. Tian; A. Reiter; G. D. Hager; T. D. Tran	UNLV-Dive
10.1109/WACV.2019.00161	2019	Action Quality Assessment Across Multiple Actions	P. Parmar; B. Morris	AQA-7
10.1109/CVPR.2019.00039	2019	What and How Well You Performed? A Multitask Learning Approach to Action Quality Assessment	P. Parmar; B. T. Morris	MTL-AQA
10.1007/978-3-030-58577-8_14	2020	An Asymmetric Modeling for Action Assessment	Gao, J.; Zheng, W.-S.; Pan, J.-H.; Gao, C.; Wang, Y.-W.; Zeng, W.; Lai, J.-H.	JIGSAWS; TASD-2; AQA-7
10.1109/ICMEW46912.2020.9106049	2020	Efficient Fitness Action Analysis Based on Spatio-Temporal Feature Encoding	J. Li; H. Cui; T. Guo; Q. Hu; Y. Shen	Fitness-AQA
10.1145/3394171.3413560	2020	Hybrid Dynamic-static Context-aware Attention Network for Action Assessment in Long Videos	Zeng, L.-A.; Hong, F.-T.; Zheng, W.-S.; Yu, Q.-Z.; Zeng, W.; Wang, Y.-W.; Lai, J.-H.	Rhythmic Gymnastics
10.3390/electronics9040568	2020	Learning effective skeletal representations on RGB video for fine-grained human action quality assessment	Lei, Q.; Zhang, H.; Du, J.; Hsiao, T.-C.; Chen, C.-C.	MIT-Skate; MIT-Dive; UNLV-Skate; UNLV-Dive; UNLV-Vault
10.1109/TransAI49837.2020.00030	2020	Skeleton-Based Detection of Abnormalities in Human Actions Using Graph Convolutional Networks	B. X. B. Yu; Y. Liu; K. C. C. Chan	UI-PRMD
10.1145/3341105.3374092	2020	Towards a data-driven method for RGB video-based hand action quality assessment in real time	Wang, T.; Jin, M.; Wang, J.; Wang, Y.; Li, M.	Origami Video

Continued on next page

Table A.1 – Continued from previous page

DOI	Year	Title	Authors	Data Source
10.1109/CVPR42600.2020.00986	2020	Uncertainty-Aware Score Distribution Learning for Action Quality Assessment	Y. Tang; Z. Ni; J. Zhou; D. Zhang; J. Lu; Y. Wu; J. Zhou	AQA-7; MTL-AQA
10.1109/TCSVT.2020.3017727	2021	Action Quality Assessment Using Siamese Network-Based Deep Metric Learning	Jain, Hiteshi; Harit, Gaurav; Sharma, Avinash	MIT-Dive; MTL-AQA; UNLV-Dive
10.1109/WACV48630.2021.00044	2021	EAGLE-Eye: Extreme-pose Action Grader using detaiL bird's-Eye view	M. Nekoui; F. O. Tito Cruz; L. Cheng	AQA-7; ExPose; MIT-Skate
10.1109/ICCV48922.2021.00782	2021	Group-aware Contrastive Regression for Action Quality Assessment	X. Yu; Y. Rao; W. Zhao; J. Lu; J. Zhou	AQA-7; MTL-AQA; JIGSAWS
10.1145/3462244.3479891	2021	Improving the Movement Synchrony Estimation with Action Quality Assessment in Children Play Therapy	Li, J.; Bhat, A.; Barmaki, R.L.	PT13
10.1016/j.knosys.2021.107388	2021	Learning and fusing multiple hidden substages for action quality assessment	Dong, L.-J.; Zhang, H.; Shi, Q.; Lei, Q.; Du, J.; Gao, S.	UNLV-Dive
10.1109/MMSP53017.2021.9733638	2021	Piano Skills Assessment	P. Parmar; J. Reddy; B. Morris	PISA
10.1109/ITME53901.2021.00048	2021	Skeleton Based Action Quality Assessment of Figure Skating Videos	H. -Y. Li; Q. Lei; H. -B. Zhang; J. -X. Du	MIT-Skate
10.1016/j.patcog.2021.108095	2021	Skeleton-based human action evaluation using graph convolutional network for monitoring Alzheimer's progression	Yu, B.X.B.; Liu, Y.; Chan, K.C.C.; Yang, Q.; Wang, X.	UI-PRMD
10.1007/s11263-021-01486-4	2021	SportsCap: Monocular 3D Human Motion Capture and Fine-Grained Understanding in Challenging Sports Videos	Chen, X.; Pang, A.; Yang, W.; Ma, Y.; Xu, L.; Yu, J.	AQA-7; Diving48; FineGym; MTL-AQA; SMART
10.1007/s11760-021-01890-w	2021	Temporal attention learning for action quality assessment in sports video	Lei, Q.; Zhang, H.; Du, J.	AQA-7
10.1145/3453892.3461624	2021	Towards Improved and Interpretable Action Quality Assessment with Self-Supervised Alignment	Roditakis, K.; Makris, A.; Argyros, A.	MTL-AQA
10.1145/3474085.3475438	2021	TSA-Net: Tube Self-Attention Network for Action Quality Assessment	Wang, S.; Yang, D.; Zhai, P.; Chen, C.; Zhang, L.	AQA-7; MTL-AQA
10.1016/j.jvcir.2021.103304	2021	What and how well you exercised? An efficient analysis framework for fitness actions	Li, J.; Hu, Q.; Guo, T.; Wang, S.; Shen, Y.	Fitness-28
10.1007/978-3-031-19772-7_25	2022	Action Quality Assessment with Temporal Parsing Transformer	Bai, Y.; Zhou, D.; Zhang, S.; Wang, J.; DIng, E.; Guan, Y.; Long, Y.; Wang, J.	AQA-7; JIGSAWS; MTL-AQA
10.3390/electronics11193051	2022	An Efficient Motion Registration Method Based on Self-Coordination and Self-Referential Normalization	Ren, Y.; Zhang, B.; Chen, J.; Guo, L.; Wang, J.	KTH

Continued on next page

Table A.1 – Continued from previous page

DOI	Year	Title	Authors	Data Source
10.1155/2022/9402195	2022	Design of a Professional Sports Competition Adjudication System Based on Data Analysis and Action Recognition Algorithm	Lin, Q.; Zou, J.	Northwestern-UCLA; MSRC-12; CAD-60
10.1007/978-3-031-19839-7_7	2022	Domain Knowledge-Informed Self-supervised Representations for Workout Form Assessment	Parmar, P.; Gharat, A.; Rhodin, H.	Fitness-AQA
10.1109/CVPR52688.2022.00296	2022	FineDiving: A Fine-grained Dataset for Procedure-aware Action Quality Assessment	J. Xu; Y. Rao; X. Yu; G. Chen; J. Zhou; J. Lu	FineDiving
10.1038/s41597-022-01188-7	2022	Functional movement screen dataset collected with two Azure Kinect depth sensors	Xing, Qing-Jun; Shen, Yuan-Yuan; Cao, Run; Zong, Shou-Xin; Zhao, Shu-Xiang; Shen, Yan-Fei	self-created
10.1007/978-3-031-04881-4_46	2022	Improving Action Quality Assessment Using Weighted Aggregation	Farabi, S.; Himel, H.; Gazzali, F.; Hasan, M.B.; Kabir, M.H.; Farazi, M.	MTL-AQA
10.1016/j.patrec.2022.04.015	2022	Learning time-aware features for action quality assessment	Zhang, Y.; Xiong, W.; Mi, S.	MTL-AQA
10.1109/CVPR52688.2022.00323	2022	Likert Scoring with Grade Decoupling for Long-term Action Assessment	A. Xu; L. -A. Zeng; W. -S. Zheng	FIS-V; JIGSAWS; Rhythmic Gymnastics
10.1007/978-3-031-19772-7_27	2022	Pairwise Contrastive Learning Network for Action Quality Assessment	Li, M.-Z.; Zhang, H.; Lei, Q.; Fan, Z.; Liu, J.; Du, J.	AQA-7; MTL-AQA
10.1109/ISCAS48785.2022.9937262	2022	RARN: A Real-Time Skeleton-based Action Recognition Network for Auxiliary Rehabilitation Therapy	M. Shen; H. Lu	RDSD
10.1109/ICIP46576.2022.9897932	2022	Representation Learning Using Rank Loss for Robust Neurosurgical Skills Evaluation	B. Baby; M. Chasmai; T. Banerjee; A. Suri; S. Banerjee; C. Arora	JIGSAWS; NETS
10.1155/2022/5430463	2022	Research on Tennis Motion Evaluation Method and System Based on Deep Learning	Chang, Huan	MSR-Action3D
10.1109/TCSVT.2022.3143549	2022	Semi-Supervised Action Quality Assessment With Self-Supervised Segment Feature Recovery	S. -J. Zhang; J. -H. Pan; J. Gao; W. -S. Zheng	MTL-AQA; Rhythmic Gymnastics
10.1016/j.jvcir.2022.103625	2022	Skeleton-based deep pose feature learning for action quality assessment on figure skating videos	Li, H.; Lei, Q.; Zhang, H.; Du, J.; Gao, S.	MIT-Skate; FIS-V
10.1007/978-3-031-18913-5_17	2022	Skeleton-Based Action Quality Assessment via Partially Connected LSTM with Triplet Losses	Wang, X.; Li, J.; Hu, H.	UMONS-TAICHI; Walking Gait
10.1109/MMSP55362.2022.9949464	2022	Tai Chi Action Quality Assessment and Visual Analysis with a Consumer RGB-D Camera	J. Li; H. Hu; Q. Xing; X. Wang; J. Li; Y. Shen	TaiChi-24
10.1145/3581783.3613774	2023	A Figure Skating Jumping Dataset for Replay-Guided Action Quality Assessment	Liu, Y.; Cheng, X.; Ikenaga, T.	RFSJ

Continued on next page

Table A.1 – Continued from previous page

DOI	Year	Title	Authors	Data Source
10.1007/978-3-031-44216-2_19	2023	A Graph Convolutional Siamese Network for the Assessment and Recognition of Physical Rehabilitation Exercises	Li, C.; Ling, X.; Xia, S.	UI-PRMD; IRDS
10.1117/12.2685368	2023	A novel blind action quality assessment based on Multi-headed GRU network and attention mechanism	Sun, W.; Hu, Y.; Zhang, B.; Chen, X.; Hao, C.; Gao, Y.	AQA-7; JIGSAWS
10.1155/2023/3649217	2023	A Novel Model for Intelligent Pull-Ups Test Based on Key Point Estimation of Human Body and Equipment	Liu, G.; Wang, J.; Zhang, Z.; Liu, Q.; Ren, Y.; Zhang, M.; Chen, S.; Bai, P.	self-created
10.1109/EIECC60864.2023.10456639	2023	A Temporal Action Evaluation Algorithm for Medical Clinical Skill Operations	Z. Wang; J. Ruan; J. Zhang; J. Yue	Thumos14; ActivityNet1.3
10.1109/TVCG.2023.3247092	2023	A Video-Based Augmented Reality System for Human-in-the-Loop Muscle Strength Assessment of Juvenile Dermatomyositis	K. Zhou; R. Cai; Y. Ma; Q. Tan; X. Wang; J. Li; H. P. H. Shum; F. W. B. Li; S. Jin; X. Liang	JDM; 3D Animation
10.1109/ICMLC58545.2023.10327994	2023	Action Quality Assessment for ASD Behaviour Evaluation	D. Zhang; D. Zhou; H. Liu	AQA-7; BEST; DREAM; EPIC-Skills; FIS-V; Infinite Grasp; JIGSAWS; MIT-Dive
10.1109/ACCESS.2023.3316009	2023	AI Trainer: Autoencoder Based Approach for Squat Analysis and Correction	M. Chariar; S. Rao; A. Irani; S. Suresh; C. S. Asha	Countix; Fitness-AQA; Kinetics 700; MM-Fit; UCF-101
10.1109/TNSRE.2023.3317411	2023	A Contrastive Learning Network for Performance Metric and Assessment of Physical Rehabilitation Exercises	L. Yao; Q. Lei; H. Zhang; J. Du; S. Gao	UI-PRMD; KIMORE
10.1007/s00521-023-09068-w	2023	Auto-encoding score distribution regression for action quality assessment	Zhang, Boyu; Chen, Jiayuan; Xu, Yinfei; Zhang, Hui; Yang, Xu; Geng, Xin	AQA-7; MTL-AQA; JIGSAWS
10.1007/s11263-022-01695-5	2023	Automatic Modelling for Interactive Action Assessment	Gao, J.; Pan, J.-H.; Zhang, S.-J.; Zheng, W.-S.	JIGSAWS; TASD-2; PaSk; AQA-7
10.1109/ACCESS.2023.3305372	2023	Contrastive Learning for Action Assessment Using Graph Convolutional Networks With Augmented Virtual Joints	C. -I. Joung; S. Byun; S. Baek	AI-Hub Fitness; Squat
10.1109/TIP.2023.3331212	2023	Fine-Grained Spatio-Temporal Parsing Network for Action Quality Assessment	K. Gedamu; Y. Ji; Y. Yang; J. Shao; H. T. Shen	FineDiving; AQA-7; MTL-AQA
10.1007/s40747-022-00892-6	2023	Gaussian guided frame sequence encoder network for action quality assessment	M.-Z. Li; Zhang, H.; Dong, L.-J.; Lei, Q.; Du, J.	AQA-7; MTL-AQA
10.1109/TCSVT.2023.3281413	2023	Hierarchical Graph Convolutional Networks for Action Quality Assessment	K. Zhou; Y. Ma; H. P. H. Shum; X. Liang	AQA-7; MTL-AQA; JIGSAWS
10.1007/s10489-023-05166-3	2023	Improving action quality assessment with across-staged temporal reasoning on imbalanced data	Lian, P.-X.; Shao, Z.-G.	FineDiving

Continued on next page

Table A.1 – Continued from previous page

DOI	Year	Title	Authors	Data Source
10.1007/s10489-022-03984-5	2023	Label-reconstruction-based pseudo-subscore learning for action quality assessment in sporting events	Zhang, H.; Dong, L.-J.; Lei, Q.; Yang, L.-J.; Du, J.	AQA-7; MTL-AQA
10.1145/3581783.3613795	2023	Localization-assisted Uncertainty Score Disentanglement Network for Action Quality Assessment	Ji, Y.; Ye, L.; Huang, H.; Mao, L.; Zhou, Y.; Gao, L.	FineFS
10.1109/CVPR52729.2023.00238	2023	LOGO: A Long-Form Video Dataset for Group Action Quality Assessment	S. Zhang; W. Dai; S. Wang; X. Shen; J. Lu; J. Zhou; Y. Tang	LOGO
10.1145/3650400.3650642	2023	Martial arts Scoring System based on U-shaped networkWushu intelligent scoring systemLearning to score Chinese Wushu	Li, M.; Tian, F.; Li, Y.	AGF-Olympics; FIS-V; Fri2023; LOGO; MTL-AQA
10.3390/systems11010021	2023	MLA-LSTM: A Local and Global Location Attention LSTM Learning Model for Scoring Figure Skating	Han, C.; Shen, F.; Chen, L.; Lian, X.; Gou, H.; Gao, H.	FIS-V; MIT-Skate
10.1145/3577190.3614117	2023	MMASD: A Multimodal Dataset for Autism Intervention Analysis	Li, J.; Chheang, V.; Kullu, P.; Brignac, E.; Guo, Z.; Bhat, A.; Barner, K.E.; Barmaki, R.L.	MMASD
10.1109/MMSP59012.2023.10337711	2023	MR-STGN: Multi-Residual Spatio Temporal Graph Network Using Attention Fusion for Patient Action Assessment	Y. Mourchid; R. Slama	UI-PRMD
10.1145/3622896.3622916	2023	Multi-Stage Action Quality Assessment Method	Liu, L.; Zhai, P.; Zheng, D.; Fang, Y.	FineDiving
10.1007/s10489-023-04613-5	2023	Multi-skeleton structures graph convolutional network for action quality assessment in long videos	Lei, Q.; Li, H.; Zhang, H.; Du, J.; Gao, S.	MIT-Skate; Rhythmic Gymnastics
10.1145/3606038.3616150	2023	Video-based Skill Assessment for Golf: Estimating Golf Handicap	Ingwersen, C.K.; Xarles, A.; Clapés, A.; Madadi, M.; Jensen, J.N.; Hannemose, M.R.; Dahl, A.B.; Escalera, S.	Golf Swing Video
10.1109/TMM.2022.3222681	2023	View-Invariant Center-of-Pressure Metrics Estimation With Monocular RGB Camera	C. Du; S. Graham; C. Depp; T. Nguyen	self-created
10.26555/ijain.v9i1.919	2023	Who danced better? Ranked tiktok dance video dataset and pairwise action quality assessment method	Hipiny, I.; Ujir, H.; Alias, A.A.; Shanat, M.; Ishak, M.K.	Ranked TikTok (CDRG-UNIMAS)
10.1109/ICAACE61206.2024.10548995	2024	Action Quality Assessment with Multi-scale Temporal Attention Mechanism	W. Wang; H. Wang; Y. Hao; Q. Wang	AQA-7; MTL-AQA
10.1109/TMM.2023.3294800	2024	Adaptive Stage-Aware Assessment Skill Transfer for Skill Determination	S. -J. Zhang; J. -H. Pan; J. Gao; W. -S. Zheng	AQA-7; BEST; EPIC-Skills; JIGSAWS
10.1007/s10489-024-05349-6	2024	Assessing action quality with semantic-sequence performance regression and densely distributed sample weighting	Huang, F.; Li, J.	UNLV-Dive; AQA-7

Continued on next page

Table A.1 – Continued from previous page

DOI	Year	Title	Authors	Data Source
10.1109/TPAMI.2024.3378753	2024	EGCN++: A New Fusion Strategy for Ensemble Learning in Skeleton-Based Rehabilitation Exercise Assessment	Yu, B.X.B.; Liu, Y.; Chan, K.C.C.; Chen, C.W.	UI-PRMD; KIMORE; EHE
10.1109/CVPR52733.2024.01386	2024	FineParser: A Fine-grained Spatio-temporal Action Parser for Human-centric Action Quality Assessment	Xu, J.; Yin, S.; Zhao, G.; Wang, Z.; Peng, Y.	FineDiving
10.1109/CVPRW63382.2024.00324	2024	FineRehab: A Multi-modality and Multi-task Dataset for Rehabilitation Analysis	Li, J.; Xue, J.; Cao, R.; Du, X.; Mo, S.; Ran, K.; Zhang, Z.	FineRehab
10.1109/CBMS61543.2024.00016	2024	GYMetricPose: A light-weight angle-based graph adaptation for action quality assessment	Gallardo, U.; Caro, F.; Hernandez, E.; Espinosa, R.; Ochoa-Ruiz, G.	Fitness-AQA
10.1109/TMM.2023.3328180	2024	Learning Semantics-Guided Representations for Scoring Figure Skating	Z. Du; D. He; X. Wang; Q. Wang	OlympicFS
10.1109/TIM.2024.3398072	2024	Learning Sparse Temporal Video Mapping for Action Quality Assessment in Floor Gymnastics	S. Zahan; G. Mubashar Hassan; A. Mian	AGF-Olympics
10.1016/j.compbiomed.2024.108382	2024	LightPRA: A Lightweight Temporal Convolutional Network for Automatic Physical Rehabilitation Exercise Assessment	Sardari, S.; Sharifzadeh, S.; Daneshkhah, A.; Loke, S.W.; Palade, V.; Duncan, M.J.; Nakisa, B.	UI-PRMD; KIMORE
10.1109/ACCESS.2024.3423462	2024	MMW-AQA: Multimodal In-the-Wild Dataset for Action Quality Assessment	T. Nagai; S. Takeda; S. Suzuki; H. Seshimo	MMW-AQA
10.1109/ICASSP48485.2024.10447069	2024	Multi-Stage Contrastive Regression for Action Quality Assessment	Q. An; M. Qi; H. Ma	FineDiving
10.1109/TIP.2024.3362135	2024	Multimodal Action Quality Assessment	L. -A. Zeng; W. -S. Zheng	FIS-V; FineDiving; Rhythmic Gymnastics
10.1109/CVPR52733.2024.01744	2024	Narrative Action Evaluation with Prompt-Guided Multimodal Interaction	Zhang, S.; Bai, S.; Chen, G.; Chen, L.; Lu, J.; Wang, J.; Tang, Y.	MTL-AQA; FineGym
10.1109/WACV57701.2024.00012	2024	PECoP: Parameter Efficient Continual Pretraining for Action Quality Assessment	A. Dadashzadeh; S. Duan; A. Whone; M. Mirmehdi	JIGSAWS; MTL-AQA; FineDiving; PD4T
10.1007/978-3-031-61678-5_6	2024	Sport Action Evaluation Based on Human Pose Estimation: A Case of Evaluating Golf Swing Action	Li, X.; Tao, R.; Wang, Y.; Ding, Y.; Luo, X.	Golf Swing Action
10.1016/j.ins.2024.120347	2024	Two-path target-aware contrastive regression for action quality assessment	Ke, X.; Xu, H.; Lin, X.; Guo, W.	MTL-AQA; FineDiving; AQA-7; JIGSAWS
10.1109/ICASSP48485.2024.10448158	2024	Which is the Better Teacher Action? A New Ranking Model and Dataset	M. Fang; X. Du; Q. Liu; Y. Zhou; Q. Liang; S. Liu	TAQR

Table A.2: Summary of datasets extrapolated from the reference study that were not included in the framework.

Dataset	Relation	Avail.	Article DOI	Original DOI	Incl.	Cit.	Orig.
CPR-Coach	AQA-ready	Public	10.1109/LSP.2023.3346207	10.1109/CVPR52733.2024.01777	No	8	7
FS1000	AQA-ready	Public	10.1109/TMM.2023.3328180	10.1609/aaai.v37i3.25392	Yes	25	30
Perturbed Workout	AQA-ready	Private	10.1145/2505323.2505330	–	No	55	–
SkatingVerse	AQA-ready	Public	10.1049/cvi2.12287	–	No	4	–
Stroke Rehabilitation	AQA-ready	Private	10.1109/CVPRW.2013.82	–	No	42	–
SU-41 Gesture	AQA-related	Private	10.1145/2505323.2505330	–	No	55	–
TGC20ReId	AQA-ready	Private	10.1007/978-3-031-06433-3_21	10.1016/j.patrec.2020.08.003	No	8	27
Average						**28.14**	**21.33**
Total						**197**	**64**

Note: *Relation*: AQA-ready, AQA-related (adaptable), Not applicable. Relation labels are consensus annotations from three annotators working independently. *Incl.*: Article included in the framework. *Cit./Orig.*: Citation counts from Google Scholar (Dec 20, 2025).

Table A.3: Summary of datasets extrapolated with the proposed framework that were not included in the reference study.

Dataset	Relation	Avail.	Article DOI	Original DOI	Incl.	Cit.	Orig.
3D Animation	AQA-ready	Private	10.1109/TVCG.2023.3247092	–	Yes	25	–
ActivityNet1.3	AQA-related	Public	10.1109/EIECC60864.2023.10456639	10.1109/CVPR.2015.7298698	No	0	3,536
AGF-Olympics	AQA-ready	Public	10.1109/TIM.2024.3398072	–	No	15	–
CAD-60	AQA-related	Public	10.1155/2022/9402195	10.1109/ICRA.2012.6224591	No	3	754
Countix	AQA-related	Public	10.1109/ACCESS.2023.3316009	10.48550/arXiv.2006.15418	Yes	31	183
Diving48	AQA-related	Public	10.1007/s11263-021-01486-4	10.1007/978-3-030-01231-1_32	No	73	224
ExPose	AQA-ready	Public	10.1109/WACV48630.2021.00044	10.1109/CVPRW50498.2020.00458	Yes	39	41
FINA09	AQA-ready	Private	10.1109/CVPRW.2017.16	10.1007/978-3-642-22822-3_27	No	283	17
FineRehab	AQA-ready	Public	10.1109/CVPRW63382.2024.00324	–	No	18	–
Fitness-AQA	AQA-ready	Private	10.1109/ACCESS.2023.3316009	10.48550/arXiv.2202.14019	Yes	31	42

Continued on next page

Table A.3 – Continued from previous page

Dataset	Relation	Avail.	Article DOI	Original DOI	Incl.	Cit.	Orig.
Golf Swing Action	AQA-related	Private	10.1007/978-3-031-61678-5_6	–	No	0	–
Golf Swing Video	AQA-related	Private	10.1145/3606038.3616150	–	Yes	7	–
Infinite Grasp	AQA-ready	Private	10.1109/ICMLC58545.2023.10327994	10.48550/arXiv.1901.02579	Yes	0	76
Kinetics 700	AQA-related	Public	10.1109/ACCESS.2023.3316009	10.48550/arXiv.1907.06987	Yes	31	681
KTH	AQA-related	Public	10.3390/electronics11193051	10.1109/ICPR.2004.1334462	No	0	5,397
MM-Fit	AQA-related	Public	10.1109/ACCESS.2023.3316009	10.1145/3432701	Yes	31	91
MSR-Action3D	AQA-ready	Public	10.1016/j.patcog.2017.12.007	10.1109/CVPRW.2010.5543273	No	128	1,936
MSRC-12	AQA-related	Public	10.1016/j.patcog.2017.12.007	10.1109/CVPR.2011.5995316	No	128	4,935
Northwestern-UCLA	AQA-related	Public	10.1155/2022/9402195	10.1109/CVPR.2011.5995729	No	3	360
Ranked TikTok (CDRG-UNIMAS)	AQA-ready	Private	10.26555/ijain.v9i1.919	–	No	7	–
RDSD	AQA-related	Private	10.1109/ISCAS48785.2022.9937262	–	No	5	–
SMART	AQA-ready	Public	10.1007/s11263-021-01486-4	–	No	73	–
TAQR	AQA-ready	Public	10.1109/ICASSP48485.2024.10448158	–	No	5	–
Thumos14	AQA-related	Public	10.1109/EIECC60864.2023.10456639	10.1016/j.cviu.2016.10.018	No	0	747
UCF-101	AQA-related	Public	10.1109/ACCESS.2023.3316009	10.48550/arXiv.1212.0402	Yes	31	8,393
Average						**38.68**	**1,713.31**
Total						**967**	**27,413**

Note: *Relation*: AQA-ready, AQA-related (adaptable), Not applicable. Relation labels are consensus annotations from three annotators working independently. *Incl.*: Article included in the reference study. *Cit./Orig.*: Citation counts from Google Scholar (Dec 20, 2025).

Appendix B

This appendix provides the detailed results of the robustness analyses. It reports discrepancies observed across repeated executions of the primary framework, the additional dataset mentions recovered through full-text RAG, and the study-level and dataset-level differences between the Gemini- and Claude-based executions.

Table B.1: Inter-run discrepancies for study inclusion and dataset extraction in the PRISMA phases.

Stage	DOI	Majority Output	Variant Output	Reason for Discrepancy
Screening (145 records)				
Inclusion/Exclusion	10.1109/CVPR52733.2024.01744	Include (3 runs)	Exclude (2 runs)	Abstract discusses general action recognition methods without unambiguous AQA relevance, causing the LLM to oscillate on the strict inclusion criteria.
	10.1016/j.patcog.2017.12.007	Include (3 runs)	Exclude (2 runs)	Abstract discusses general action recognition methods without unambiguous AQA relevance, causing the LLM to oscillate on the strict inclusion criteria.
	10.1109/SSCI.2017.8285270	Include (3 runs)	Exclude (2 runs)	Abstract discusses general action recognition methods without unambiguous AQA relevance, causing the LLM to oscillate on the strict inclusion criteria.
Dataset Extraction	10.1109/CVPR.2019.00039	[Empty] (4 runs)	MTL-AQA (1 run)	Dataset extracted in a single run due to an implicit or vague mention in the abstract.
	10.1109/CVPR52733.2024.01744	MTL-AQA; FineGym (3 runs)	[Empty] (2 runs)	Extraction discrepancy is directly propagated from the upstream screening decision (extracted only when included).
	10.1016/j.patcog.2017.12.007	MSR-Action3D; MSRC-12 (3 runs)	[Empty] (2 runs)	Extraction discrepancy is directly propagated from the upstream screening decision (extracted only when included).
	10.1109/SSCI.2017.8285270	self-created (3 runs)	[Empty] (2 runs)	Extraction discrepancy is directly propagated from the upstream screening decision (extracted only when included).
Eligibility (21 records)				
Dataset Extraction	10.1007/s11263-021-01486-4	AQA-7; Diving48; FineGym; MTL-AQA; SMART (4 runs)	MTL-AQA; SMART (1 run)	Variance is confined to secondary or baseline datasets, while primary evaluation datasets remain fully consistent.
Dataset Extraction	10.1109/TCSVT.2020.3017727	MIT-Dive; MTL-AQA; UNLV-Dive (4 runs)	MTL-AQA (1 run)	Variance is confined to secondary or baseline datasets, while primary evaluation datasets remain fully consistent.

Table B.2: Articles for which full-text RAG-based extraction recovered additional dataset mentions within the screening-positive subset.

DOI	Dataset(s) at screening	Additional dataset(s) via full-text RAG	Already in corpus
10.1007/978-3-031-19839-7_7	Fitness-AQA	Waseda Backsquat	No
10.1109/CBMS61543.2024.00016	Fitness-AQA	AQA-7	Yes
10.1109/TMM.2023.3328180	OlympicFS	FineDiving	Yes
10.1145/3622896.3622916	FineDiving	AQA-7; MTL-AQA	Yes
10.1109/TCSVT.2022.3143549	MTL-AQA; Rhythmic Gymnastics	JIGSAWS	Yes
10.1016/j.jvcir.2022.103625	FIS-V; MIT-Skate	ExPose	Yes
10.1109/MMSP55362.2022.9949464	TaiChi-24	FineDiving	Yes
10.1109/CVPR42600.2020.00986	AQA-7; MTL-AQA	JIGSAWS	Yes
10.1109/CVPR.2019.00039	MTL-AQA	MIT-Dive; UNLV-Dive	Yes
10.1109/ICASSP48485.2024.10448158	TAQR	JIGSAWS	Yes

Note: "Already in corpus" indicates whether the additional dataset had already been extracted from another included article. Waseda Backsquat was the only newly recovered dataset and was confirmed by all annotators as AQA-ready.

Table B.3: Study-level disagreements between the Gemini- and Claude-based executions, their origin in the automated pipeline, and their presence in the reference study.

Article DOI	Origin of Discrepancy	In Reference Study
Studies included only by Gemini		
10.1007/978-3-030-00767-6_12	Eligibility	Yes
10.1007/s10489-022-03984-5	Eligibility	Yes
10.1109/ACCESS.2023.3316009	Eligibility	Yes
10.1109/ICMEW46912.2020.9106049	Eligibility	No
10.1109/TMM.2023.3294800	Eligibility	Yes
10.1109/TransAI49837.2020.00030	Screening	Yes
10.1145/3606038.3616150	Eligibility	Yes
Studies included only by Claude		
10.1007/978-3-030-20876-9_10	Eligibility	Yes
10.1109/ICCT.2015.7399879	Eligibility	Yes
10.1109/ICIP42928.2021.9506257	Eligibility	Yes

Note: The table reports studies included in the final corpus by only one model execution. "Origin of Discrepancy" identifies the first stage at which the final inclusion paths diverged. "Screening" indicates that the study was retained after title-and-abstract assessment by only one model. "Eligibility" indicates that the study passed screening in both executions but obtained a different final outcome during the subsequent dataset-based eligibility procedure. "In Reference Study" indicates whether the article belongs to the benchmark set of 94 studies. Gemini and Claude included 84 studies in common.

Table B.4: Dataset-level disagreements between the Gemini- and Claude-based executions and their presence in the reference study.

Dataset	Source Article DOI	In Reference Study
Extracted only by Gemini		
3D Animation	10.1109/TVCG.2023.3247092	No
BEST	10.1109/ICMLC58545.2023.10327994	Yes
Countix	10.1109/ACCESS.2023.3316009	No
Diving48	10.1007/s11263-021-01486-4	No
EPIC-Skills	10.1109/ICMLC58545.2023.10327994	Yes
FINA09	10.1109/CVPRW.2017.16	No
Golf Swing Action	10.1007/978-3-031-61678-5_6	No
Infinite Grasp	10.1109/ICMLC58545.2023.10327994	No
Kinetics 700	10.1109/ACCESS.2023.3316009	No
MM-Fit	10.1109/ACCESS.2023.3316009	No
Ranked TikTok (CDRG-UNIMAS)	10.26555/ijain.v9i1.919	No
Extracted only by Claude		
HMDB51	10.1109/CVPRW.2017.16	No
MPII	10.1109/WACV48630.2021.00044	No
Sports-1M	10.1109/CVPRW.2017.16	No

Note: The table reports datasets extracted by only one model execution after applying the same harmonization procedure. "Source Article DOI" denotes the article from which the dataset name was extracted, while "In Reference Study" indicates whether the dataset belongs to the set of 47 datasets reported in the reference review. Gemini and Claude shared 54 datasets; Gemini uniquely extracted 11 and Claude 3. Dataset-level differences do not necessarily imply study-level disagreements.

Table B.5: Computational performance and estimated costs across pipeline stages for the supplementary Claude-based execution.

Stage	Model	Calls	In-Tokens	Out-Tokens	Time (s)	Cost (USD)
Screening (145 records)						
Screening	Sonnet 4.5	10	52,903	9,435	200.6	0.3002
Harmonization	Opus 4.5	1	2,856	1,707	16.6	0.0570
Subtotal	–	**11**	**55,759**	**11,142**	**217.2**	**0.3572**
Eligibility (29 records)						
Vector Indexing	Embedding 2	1	389,806	–	72.3	0.0780
Extraction (RAG)	Sonnet 4.5	28	73,393	549	154.5	0.2284
Harmonization	Opus 4.5	1	957	303	4.7	0.0124
Subtotal	–	**30**	**464,156**	**852**	**231.5**	**0.3187**
Total	–	**41**	**519,915**	**11,994**	**448.7**	**0.6759**

Note: Values refer to a single supplementary Claude-based execution and are reported as exact measurements.