LLM-Powered Educational Conversational Agent for Open Educational Resources

Renan Zafalon da Silva[1], Paulo Cesar Ramos Pinho[1], Ulian Gabriel Alff Ramires[1], Raymundo Carlos Machado Ferreira Filho[2], Tiago Thompen Primo[1]

[1]Programa de Pós-Graduação em Ciência da Computação
Universidade Federal de Pelotas (UFPel) Pelotas - RS - Brazil

[2]Instituto Federal Sul-rio-grandense Câmpus Pelotas Visconde da Graça
Pelotas - RS - Brazil

{renan.zafalon, paulo.pinho, ugaramires, tiago.primo}@inf.ufpel.edu.br,
raymundofilho@ifsul.edu.br

***Abstract. Research Context**: Educational repositories gather diverse Open Educational Resources (OER), yet sparse metadata and inconsistent terminology reduce findability. Large Language Models (LLMs) with retrieval-augmented generation (RAG) can bridge vocabulary gaps by capturing semantic similarity, thereby improving recall and user experience. **Scientific and/or Practical Problem**: The national OER repository (ProEdu) depends on a solely lexical engine. This dependence creates difficulties in handling synonyms, paraphrases, and domain shifts, resulting in suboptimal recall and inconsistent rankings. **Proposed Solution and/or Analysis**: We develop a prototype of an Educational Conversational Agent (ECA) that integrates retrieval and response generation. Three pipelines are evaluated: ProEdu, which employs a lexical approach; a field-weighted Elasticsearch (ES); and a semantic RAG system utilizing Sentence-Transformers (all-MiniLM-L6-v2) embeddings with a FAISS (Facebook AI Similarity Search) index and Llama for text generation, plus a lightweight reranking mechanism. **Related IS Theory**: We assert that AI-enhanced repositories diminish search obstacles and assist educators in effectively identifying suitable materials. Furthermore, the conversational interface alleviates the cognitive load by providing verified sources within context. **Research Method**: A comparative assessment involved 22 interdisciplinary prompts in ten domains. For each prompt, we established gold-standard datasets, formulated standardized queries, and calculated precision, recall, and F1-score. **Summary of Results**: The Llama/FAISS pipeline achieves the best coverage-relevance balance driven by high recall. ES attains a similar F1 through higher precision but lower recall. ProEdu performs poorly in F1. Error analysis shows semantic retrieval excels in cross-vocabulary matches and multi-facet intents. **Contributions and Impact to IS area**: We deliver a replicable benchmark for large-scale OER search (prompts, metrics, code) and a pragmatic architecture combining semantic RAG and ES to balance recall and precision on cost-efficient infrastructure. Prompt templates and evaluation scripts support adoption.*

1. Research Context

Educational repositories serve as essential information systems in academia and industry, facilitating the storage, organization, and dissemination of learning materials. These platforms often contain Open Educational Resources (OER) and support a variety of teaching and research activities by providing centralized access to diverse educational content. Efficient search capabilities in such repositories are crucial to maximize their value and usability. For example, the repository studied in this work (ProEdu) has more than 20 million downloads, underscoring the importance of effective retrieval mechanisms to fully leverage its content. ProEdu is an online repository that currently indexes more than 1,800 OER focused on professional and technological education. It offers a wide range of educational materials—including courses, thematic notebooks, books, animations, audio resources, videos, images, web pages, and multiple domains of knowledge. These materials are designed to support teaching, learning, and educational management processes [1].

ProEdu implements a stringent curation protocol orchestrated by a multidisciplinary technical team to determine the quality of digital educational resources. The procedural workflow involves submission accompanied by the required metadata, including title, date, and Capes knowledge area, where authors or managers bear the responsibility of ensuring authorship, originality, and licensing. The evaluation phase assesses the educational efficacy, scientific relevance, and adherence to accessibility standards, facilitated by standardized taxonomies and instructional design frameworks. Upon completion of the revisions and validations, the approved resources are incorporated into the collection, receiving compliance certification. Official manuals and terms articulate the roles and responsibilities pertinent to each stage of the procedure.

However, many educational repositories rely on a simple lexical search, which matches queries to documents based on exact word overlap. This approach struggles with linguistic variations such as synonyms and paraphrases, causing relevant resources to be missed when query terms do not exactly match the indexed text. As a result, the user's ability to discover valuable OER is hindered, reducing the utility of the repository. Recent advances in search technology, including semantic similarity models and the incorporation of generative language models, have demonstrated the potential to produce more relevant search results. However, most repositories remain restricted to conventional lexical methodologies [Mićunović et al. 2023].

To address the identified gap, we performed an evaluation of three search strategies within an educational repository: a keyword-based lexical baseline, a field-weighted approach utilizing Elasticsearch (ES), and a semantic framework that incorporates vector embeddings alongside a generative model. From a theoretical perspective within Information Systems, this undertaking applies Design Science Research (DSR) in the design and evaluation of the Educational Conversational Agent (ECA) artifact. This process meticulously follows the six-step methodology of problem identification, objective formulation, design and development, demonstration, evaluation, and communication to guaranty practical relevance as well as scientific rigor [Gregor and Hevner 2013]. The research inquiry investigates the variations among these approaches in terms of their effectiveness in retrieving relevant OER. Through empirical evaluation, we examine whether semantic and generative methodologies improve findability compared to traditional search techniques [Li et al. 2025].

This research makes a substantial contribution to the field of Information Systems.

[1] `https://proedu.rnp.br`

Firstly, empirically illustrates that semantic search techniques, including embeddings and generative models, significantly improve the discoverability of educational resources in comparison to traditional lexical search methods. Secondly, the study designs and implements a prototype search architecture that integrates these advanced methodologies within an educational repository. Third, by evaluating these approaches on a large-scale OER repository, the research offers practical insights to enhance search functionality in educational information systems.

The structure of this article is delineated as follows: Section 1 delineates the Research Context; Section 2 articulates the Scientific and Practical Problem; Section 3 elaborates on the Proposed Solution; Section 4 positions the study within the context of Related IS Theory, incorporating the literature review therein; Section 5 elucidates the research methodology, including data set, prompts, metrics, and procedures; Section 6 presents the Results; and Section 7 synthesizes the contributions and impact to the IS area, encompassing limitations and prospective future research avenues.

2. Scientific and Practical Problem

The ProEdu search engine operates purely lexically, thus constraining synonym and variant matching and lacking contextual understanding, which subsequently diminishes the efficacy of recall and ranking. Although initial studies suggest improvements in coverage and relevance through the use of Large Language Models (LLM)-powered semantic search [Gheewala et al. 2025], there is a paucity of empirical evidence on the influence of these systems in large-scale and multidisciplinary educational contexts. Consequently, this study seeks to investigate the following research question. To what extent does an Retrieval-Augmented Generation (RAG) pipeline that uses Llama [Touvron et al. 2023] and Facebook AI Similarity Search (FAISS) [Douze et al. 2025] surpass lexical methodologies, including the existing ProEdu engine, with regard to precision and recall?

The investigation begins with a technological characterization that elucidates the architectural, algorithmic, and computational prerequisites that distinguish the three search strategies. Subsequently, a performance evaluation is performed to quantify improvements in precision, recall, F1 score, and collection coverage. The advantages and limitations are examined later.

2.1. People, Process and Technologies

This dimension refers to the use of the platform by educators and learners. Educators formulate instructional content by accessing relevant OER that correspond to pedagogical goals, thereby minimizing the time expended on filtering non-essential information and facilitating the selection of appropriate materials. Similarly, learners exploit curated repository resources for research purposes. The processes dimension encompasses workflows for the curation of content and the mediation of search activities.

All educational resources within the repository are contributed and reviewed by authorized educational personnel, ensuring the provision of high-quality, meticulously curated metadata. The technological dimension integrates advanced artificial intelligence models and search mechanisms. At its core, the system operates on an RAG framework. Semantic RAG system utilizing all MiniLM-L6-v2 embeddings with a FAISS index and Llama for text generation, complemented by two auxiliary search engines [Face 2021].

Teachers and students engage with and access content through established search and curation workflows, driven by cutting-edge AI methodologies.

3. Proposed Solution and/or Analysis

Generative AI (GenAI) synthesizes new forms of data, including text, images, and music, through the application of machine learning models that are trained on large datasets [Golda et al. 2024]. GenAI utilizes LLMs to identify patterns and produce content that is appropriate to the given context. This model possesses multilingual capabilities and improved alignment with instructional guidelines and reasoning processes, making it particularly suitable for educational applications. Facilitates the precise, context-sensitive development and systematic organization of pedagogical resources [Touvron et al. 2024]. This work proposes an educational ECA, built with an RAG architecture. In the following section, a succinct exegesis of the technical components that underlie the RAG framework is presented. This includes the integration of Llama with FAISS indexing, the configuration of ES, the operational dynamics of the RAG workflow, and considerations pertinent to information retrieval within educational repositories.

3.1. Llama and Faiss

FAISS is an open-source library intended to enable efficient and high-performance nearest-neighbor search within extensive collections of embeddings. It offers support for both exact and approximate indexing on both CPU and GPU platforms. In the current set-up, sentence embeddings are subjected to L2 normalization, with the use of the IndexFlatL2 index, an exact indexing technique noted for achieving robust recall in medium-scale corpora. Llama, an open source language model developed by Meta AI, is based on transformer architecture and is engineered to generate coherent text and perform advanced tasks in natural language understanding and generation. [Douze et al. 2025].

3.2. Elasticsearch

ES serves as a distributed platform for information retrieval and analytics, built on the foundation of Apache Lucene. Each index is divided into primary shards and replicas, which facilitates horizontal scalability, load distribution, and high availability [2]. The engine employs Okapi BM25 as its default ranking model; however, its modular architecture supports extensions such as function score queries, learning-to-rank algorithms, and k nearest neighbors (k-NN) vector search. This allows for the creation of hybrid configurations that integrate both lexical and semantic matching. Queries, ingestion pipelines, and aggregations are accessible through a RESTful API, thereby improving integration with big data workflows and real-time analysis and generation of RAG systems [Gormley and Tong 2015].

3.3. Information Retrieval in Educational Repositories

Our RAG system integrates ES and search functionalities with Llama/FAISS vector retrieval to enhance literal matching alongside semantic recall. Educational documents are subjected to a process comprising text extraction, chunking, and embedding indexing. Items retrieved are systematically aligned with educational taxonomies through metadata-based filtering. For a visual representation of our proposed solution, refer to Figure 1.

[2] https://www.elastic.co/docs/get-started

In Figure 1, the workflow originates at the Educational Repository (1), where source documents are ingested by the Document Processing and Indexing module (2). This module consists of a text extraction stage (3), followed by the segmentation of the content into retrievable units ("chunking") (4), and then each resulting segment is transformed into a dense vector representation (embedding) (5) and stored in an FAISS vector index (6), allowing retrieval based on semantic similarity. The diagram employs a standardized color-coding scheme to classify functional modules and enhance the visual organization of the discrete workflow stages.

During inference, the user query (7) is processed by the Query Processing module (8), which employs a hybrid retrieval strategy that integrates lexical retrieval through Elasticsearch (9) with vector-based retrieval through Llama + FAISS (10). The contextual information retrieved is subsequently used to condition the response generation process (11), producing an evidence-based answer for the ECA.

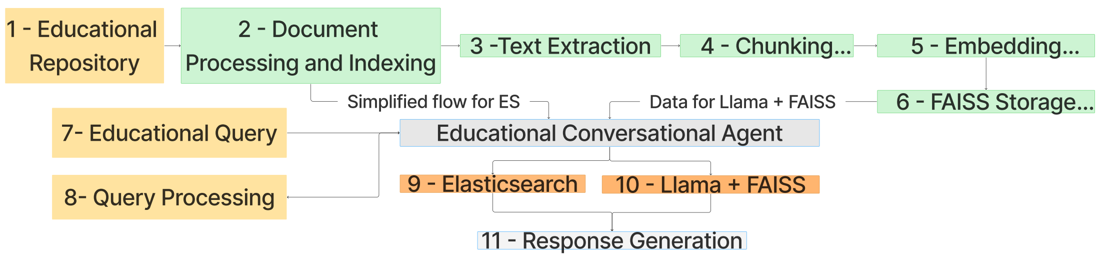

Figure 1. Diagram of the proposed solution

4. Related IS Theory

The GranDSI-BR e-book posits that research within the domain of Information Systems and associated technologies should be directed towards addressing real-world societal challenges, thereby yielding advancements with significant social utility. This premise is in direct alignment with our initiative to improve search capabilities in ProEdu through the application of generative AI.

The initiative democratizes information access by integrating LLMs with contemporary and reliable data sources, thus enabling citizens to acquire natural and relevant responses to complex inquiries. This contextualized access fosters inclusivity and civic participation, contributing to the evolution of a better informed and integrated society.

Within the educational domain, the system functions as an adaptive and collaborative platform. The semantic retrieval mechanisms used within the system are instrumental in identifying and organizing resources of pedagogical relevance, thereby directing the LLM to produce outputs aligned with the curriculum and responsive to the requirements of users. As a result, the system enhances the discovery and curation of educational knowledge while promoting improved teaching-learning practices and offering personalized support [Boscarioli et al. 2017].

4.1. Related Work

This subsection conducts a comprehensive examination of GenAI methodologies in the realm of semantic search and educational recommendation systems. It analyzes their impacts on personalization and accuracy, delineates associated benefits and risks such as explainability, bias, and privacy, and explores various presentation modalities, including ranked lists, justifications, and conversational user interfaces. In addition, it considers deployment environments and data paradigms, including K–12, higher education, massive open online courses (MOOC), and corporate settings, in addition to content repositories and curation. In addition, it evaluates accessibility practices and constraints pertinent to offline utilization.

4.1.1. Systematic Review Protocol

The Systematic Literature Review (SLR) was conducted in accordance with the protocol proposed by [Kitchenham 2004]. The search string employed was: ("educational recommendation systems" OR "OER search") AND ("LLM" OR "generative AI") AND ("retrieval"). Inclusion criteria encompassed primary studies published between 2020 and 2025 that explicitly addressed AI-driven mechanisms for educational resource discovery. Publications published before 2020, derived from closed repositories, or failing to demonstrate a substantive and effective application of artificial intelligence techniques were excluded from consideration.

The term "Recommendation Systems" was included in addition to "Search" because, within the context of OER, the discovery process typically begins with an explicit user query but frequently culminates in the generation of personalized recommendations conditioned on teacher-specific profiles. The complete table with articles is in the following location [3] and references [4]. The methodology of this study involves a comprehensive search for the largest computer science digital libraries (Table 1).

4.1.2. Research Questions

To guide this study, we formulate eight research questions as follows. RQ1: What GenAI techniques are used in educational recommendation systems? RQ2: How do these techniques influence personalization and accuracy? RQ3: What additional advantages and risks (such as explainability, bias, and privacy) emerge? RQ4: In what formats are recommendations presented (such as ranked lists, justifications, or conversational user interfaces)? RQ5: In what educational contexts (such as K-12, higher education, MOOCs, or corporate environments) and data scenarios are these systems implemented? RQ6: What content repositories are utilized and how is the content curated or reused? RQ7: What accessibility features are incorporated and evaluated (for example, WCAG, UDL)? RQ8: Is offline usage supported by these systems and under what constraints with respect to devices, privacy, and bandwidth?

4.1.3. Results and Discussion in Systematic Review

Table 2 presents a systematic review that comprehensively synthesizes the principal findings of this study. Due to the limited number of studies addressing RQ7 and RQ8, these research

[3] https://osf.io/7dkz6?view_only=06b075f8309840a8b7ce812df7ecb53a
[4] https://osf.io/a29sk?view_only=06b075f8309840a8b7ce812df7ecb53a

Table 1. Search results across repositories

Repository	Initial Count	Selected Articles
IEEE	47	4
ScienceDirect	25	2
SpringerLink	44	1
ACM Digital Library	3 557	1
Google Scholar	17 000	9
MDPI	4	3
SBC Open Library	2	1

Table 2. Main related works and answers to the research questions

References	RQ1	RQ2	RQ3	RQ4	RQ5	RQ6
[Dutta et al. 2025]	LLM + mBERT	Higher accuracy	Personalized learning paths	Interview bot; Course list	Continuing education	✓
[Dehbozorgi et al. 2024]	LLM + RAG	Mentoring	Suggest pedagogical patterns	Textual guidance	Instruction support	✓
[Bucchiarone et al. 2024]	LLM (PolyGloT) + OER	Cohesive learning plans	Automatic OER curation	Editable dashboard	Educator planning	✓
[Pesovski et al. 2024]	ChatGPT	Personalized feedback	Adaptive quizzes	Content generation	Software engineering courses	✗
[Son et al. 2024]	Generative model	Dynamic content	Tutor + recommender	Proactive LMS assistant	Distance education	✓

questions are subsequently discussed using a descriptive narrative approach. In the case of RQ6, the table indicates, by means of a check mark, those instances in which repositories have been positively identified as utilized.

[Dutta et al. 2025], a multilingual course recommendation system, utilizes a DigComp interface chatbot to collect student profiles, which later employs a hybrid mBERT + LLM model to propose courses and learning sequences. This system integrates LLM with requirements engineering but is specifically aimed at online courses rather than OER repositories [Dehbozorgi et al. 2024]. A pedagogical advocate uses RAG to extract design patterns from a knowledge base and employs an LLM to provide recommendations for instructional practice, although it does not specifically emphasize OER.

[Bucchiarone et al. 2024], the PolyGloT system facilitates educators in formulating lesson plans using OER and generative artificial intelligence, performing the automatic curation of OER. The primary focus is on the development of lesson plans rather than on the quantitative evaluation of retrieval. [Pesovski et al. 2024], GPT-4 is embedded in a learning management system to program courses to produce a variety of instructional materials through the use of personas and to assess student participation. The emphasis is on the creative generation of content and qualitative assessment as opposed to retrieval.

Recent studies conducted by researchers from the Global South highlight the imperative need for the creation of educational information systems that are attuned to specific sociocultural contexts. As illustrated by [da Silva and Cazella 2020], while educational repositories aggregate content from diverse cultural backgrounds, there is a need for culturally tailored recommendations to prevent student alienation.

Research on educational recommenders focuses predominantly on LLM-based pipelines (RQ1), while accessibility and offline functionality remain underexplored. In all studies, LLM improves personalization and accuracy and allows the generation of study plans, tutoring assistants, and OER curation (RQ2–RQ3). Presentation practices favor conversational interfaces, chatbots, and virtual assistants, for more natural interaction (RQ4). In higher, continuing, large-scale on-line and distance education, digital content-rich ecosystems are common, yet only one study addresses disability-oriented accessibility (RQ5–RQ6). Five studies leverage existing repositories (for example, OER corpora, institutional catalogs, and digital libraries), while offline use is seldom examined (RQ7–RQ8).

4.1.4. Originality and Contribution

This study seeks to address existing research gaps by, in contrast to [Dutta et al. 2025], using an actual OER repository rather than focusing only on courses and performing a scale-based evaluation using traditional information retrieval metrics, namely precision, recall, and F1 score. Furthermore, in contrast to [Dehbozorgi et al. 2024], this research emphasizes tangible OER items and systematically compares retrieval strategies, including lexical analysis, ES, and embeddings, within a well-defined protocol that includes verifiable sources provided.

In contrast to [Bucchiarone et al. 2024], which focuses on the generation of plans, our research involves a repository-level quantitative evaluation of retrieval processes, alongside a characterization of the corpus to provide context with regard to performance. Additionally, while [Pesovski et al. 2024] emphasizes generated content and qualitative metrics, we present a reproducible RAG architecture, encompassing code, prompts, and metrics, specifically for OER retrieval using Llama+FAISS integrated with a provenance-oriented lexical and semantic hybrid approach.

5. Research Method

This study provides a comparative evaluation of three retrieval methodologies implemented within an educational repository: a production grade lexical engine (ProEdu using BM25), an optimized ES baseline, and a semantic RAG pipeline incorporating Llama with FAISS. The subsequent sections explain the dataset and gold standard, the query set, the evaluation metrics, and the statistical methodologies. Traditionally based on similarity-based matching, the field of information retrieval is currently being transformed by LLM-driven generative techniques, a paradigm that is garnering increasing scholarly interest [Li et al. 2025].

5.1. Validation Metrics

The quality of the search strategies is validated using three standard metrics to assess the retrieval of relevant educational resources [Sakai 2021]: Precision, recall, and F1 score.

Precision measures the proportion of recovered items that are actually relevant, as shown in equation 1.

$$\text{Precision} = \frac{\text{Relevant retrieved items}}{\text{Total retrieved items}} \tag{1}$$

Recall measures the proportion of all relevant items that were retrieved by the system, as shown in equation 2.

$$\text{Recall} = \frac{\text{Relevant retrieved items}}{\text{Total relevant items}} \tag{2}$$

The *F1-Score* is the harmonic mean of Precision and Recall, providing a single balanced metric (equation 3).

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

5.2. Experimental Methodology

The educational materials from the ProEdu repository were indexed in a vector database. This procedure enabled the assessment of semantic search capabilities by integrating the Llama model with the FAISS library. Subsequently, a local search engine utilizing ES was deployed. This framework was selected for managing text search within the educational repository due to its robustness, scalability, and high performance as an open-source solution, as well as its widespread adoption in information retrieval systems [Zmaranda et al. 2021].

In addition, the production search engine implemented in the repository under investigation, which integrates Apache Solr, was evaluated. The performance of the three approaches was quantitatively assessed using standard information retrieval metrics, namely precision, recall, and F1-score.

To assess the metrics, a series of 22 interdisciplinary prompts were developed to facilitate the retrieval of educational objects across four principal domains: Computer Science, Biological Sciences, Physical and Mathematical Sciences, and Humanities. The distribution of these prompts by domain is shown in Table 3. Each prompt was thematically defined through a reverse mapping of the ProEdu collection to discern key content. These prompts were constructed according to the methodology elucidated by [Li and Klabjan 2024], whose guide provides contemporary guidelines specifically tailored for the educational domain, thus ensuring alignment with research objectives.

The design of prompts aims to achieve thorough domain coverage while keeping the number of intents per sector to a minimum, accompanied by precise and verifiable constraints such as temporal and subject filters articulated using repository-specific terminology. It utilizes a standardized output schema (title, creator, date, type, identifier), which facilitates automatic parsing and alignment with learning resource standards. By conceptualizing prompts as specifications, reproducibility is ensured through adherence to established gold standards and conventional information retrieval metrics. This approach eliminates the need for modifications to the downstream components, thus preserving future compatibility with the learning object profiles.

The experiments were conducted using the GroqCloud API for the deployment and execution of the Llama model. For the purpose of generating vector embeddings within the FAISS database, preliminary data extracted from the ProEdu repository were utilized to support the development of this scholarly application. The dataset was supplied as an SQL dump comprising all metadata associated with the educational resources hosted in the ProEdu repository. Following the extraction of the relevant metadata, the data were transformed into a JSON file, as illustrated below in listing 1. The corresponding ground truth is now explicitly provided in [5].

5.2.1. Explanation of Prompt Creation

This section elucidates the reverse engineering process involved in the creation of prompts. The approach used aligns with the concept of reverse prompt engineering, which consists of analyzing the desired output to infer the most effective prompt structures. This methodology allows for the systematic construction of prompts by analyzing existing metadata in the repository, identifying structural and semantic patterns that guide the creation of optimized prompts [Li and Klabjan 2024].


```
[
  {
    "prompt": "Retrieve all materials in computer science, covering algorithms,
    programming in various languages, software engineering, data structures, and
    computer architecture. Provide the title, authors, publication date, material type,
    and access link.",
    "items": [
      {
        "item_id": "138f0e9c-a1da-4b61-b1e2-b5e8250c42cb",
        "title": "Operating Systems",
        "authors": "Double Blind",
        "date_available": "2016-08-11T12:40:59Z",
        "type": "final",
        "link": "Double Blind",
        "subjects": [
          "Operating systems (Computers)",
          "Computing systems"
        ]
      }
    ]
  }
]
```


Listing 1. Ground truth json (passage)

For example, the prompt P16: search for resources addressing: mapping of natural resources (digital cartography, geoprocessing, GIS, remote sensing); environmental impact analysis (EIA/RIMA, mitigation); territorial planning and land use management (ecological-economic zoning, soil management, urban-regional planning). This prompt has been specifically designed to test several critical capabilities of retrieval systems, following established practices for the evaluation of IR systems [Li and Klabjan 2024]. Firstly, its multi-thematic structure ('mapping', 'impact analysis', 'planning') assesses the systems' ability to handle complex and multifaceted information needs. Secondly, it includes a mixture of general terms ('territorial planning') and highly specific technical acronyms (GIS', 'EIA/RIMA'), which directly challenges the lexical limitations of the ProEdu engine and tests the semantic understanding of the Llama/FAISS pipeline.

[5] https://osf.io/qhdnb?view_only=06b075f8309840a8b7ce812df7ecb53a

5.2.2. Elasticsearch Configuration and Experimental Procedure

The methodological framework is composed of several distinct phases: initially, a singular ES index is instantiated. Textual fields undergo analytical processing tailored to the Portuguese language, which incorporates tokenization, stopword removal, and stemming, while identifier fields are preserved as exact terms. The gold standard dataset is meticulously organized, with each entry supplemented by its source prompt to ensure traceability.

The index undergoes both the creation and validation processes, with data ingested in batches characterized by predictable identifiers, complemented by systematic logging of both successful and unsuccessful operations. Queries encompassing multiple fields employ the AND operator, with specific field weights allocated (prioritizing the title, followed by authors/subjects), while ranking is determined by the BM25 algorithm [Kamphuis et al. 2020]. Figure 2 details the ES configuration and our experimental procedure.

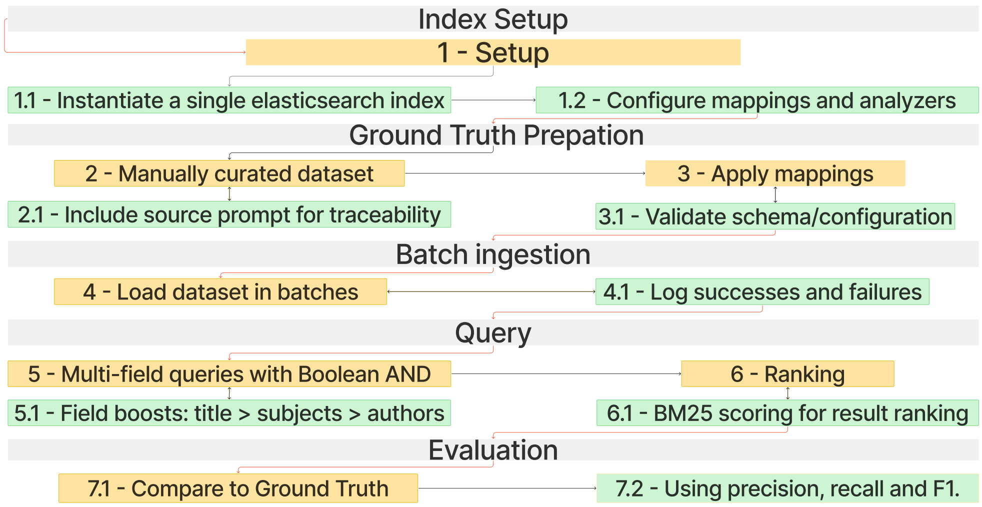

Figure 2. Elasticsearch configuration

Comprehensive empirical studies indicate that BM25 consistently surpasses traditional lexical baselines and even exceeds the performance of LLM-based dense retrieval techniques in benchmarks involving long-context and real-world retrieval scenarios. BM25 continues to be an integral part of information retrieval, despite the emergence of pre-trained LLM [Li et al. 2024].

Retrieval quality was evaluated using precision, recall, and F1 against a manually curated gold standard, ensuring thorough validation. In the learning-object search experiment, ES was configured as a strict lexical engine: all query terms had to appear, with field weights favoring titles.

This setup prioritizes precision by requiring full term overlap, filtering out partially relevant items and mostly returning highly matching documents. The AND operator yields tighter matches and fewer results, increasing precision while reducing recall. Even as a lexical engine, ES improved average precision by limiting results to the exact intersection

Table 3. Prompt distribution by area (p1-p22)

Area	Prompt Descriptions (P1–P22)
Computer Science	P1: Retrieve all materials in computer science, covering algorithms, programming in various languages, software engineering, data structures, and computer architecture. P2: Retrieve all materials in computer networks, encompassing communication protocols, network topologies, routing, network security, and infrastructure equipment. P3: Search for computing content within informatics, requesting title, authors, publication date, material type, and access link. P4: Filter items on computer networks with emphasis on router configuration, intrusion detection systems, or QoS; display title, publication date, authors, material type, and access link.
Biological Sciences	P5: Retrieve all materials in biology, including genetics, evolution, ecology, physiology, or laboratory research. P6: Restrict results to documents on molecular genetics and biotechnology published from 2015 onward, containing terms such as DNA sequencing or genetically modified organisms; present title, authors, and link.
Mathematics	P7: Retrieve all materials in mathematics, addressing algebra, geometry, calculus, statistics, or probability across different education levels. P8: Filter materials focused on problem solving, function analysis, or mathematical logic for exact sciences courses; request title, authors, publication date, and access link.
History	P9: Retrieve all contents in history.
English	P10: Retrieve all materials in English. P11: Filter English-learning content aimed at improving fluency, pronunciation, and idiomatic expressions; request title, authors, publication date, material type, and access link.
Environmental Management	P12: Retrieve all materials in environmental management, including sustainability, legislation, climate change, and case studies. P13: Filter items on applied ecology and socio-environmental responsibility, highlighting sustainability tools, environmental certifications, or biodiversity conservation; present title, authors, date, material type, and link.
Geography	P14: Retrieve all materials in geography, covering climate, landforms, vegetation, and urbanization. P15: Focus on records related to tourism, hospitality, and events; present title, publication date, authors, and access link. P16: Search for resources addressing: mapping of natural resources (digital cartography, geoprocessing, GIS, remote sensing); environmental impact analysis (EIA/RIMA, mitigation); territorial planning and land-use management (ecological–economic zoning, soil management, urban–regional planning).
Portuguese	P17: Retrieve all materials in Portuguese, including normative grammar, academic writing, and reading comprehension. P18: Filter content dedicated to the Portuguese language (study and teaching); request title, authors, publication date, and link, prioritizing academic works or book chapters.
Chemistry	P19: Retrieve all materials in chemistry, including chemical reactions, atomic structure, periodic table, or stoichiometry. P20: Restrict search to content on kinetics, chemical equilibrium, organic and inorganic chemistry, validated by experts; display title, authors, publication date, and access method.
Physics	P21: Retrieve all materials in physics, addressing mechanics, thermodynamics, electromagnetism, optics, or modern physics. P22: Filter records on physics–electromagnetism; present title, publication date, authors, material type, and access link.

of query terms and by boosting documents where the most important field, titles, also appeared among the metadata keywords of the resources.

5.2.3. Repository ProEdu Configuration and Experimental Procedure

The experimental configuration for the ProEdu repository began with defining the search prompts Table 3, engineered to maximize thematic coverage and retrieve a comprehensive set of items with key metadata (title, publication date, authorship). Using the implemented ECA interface (see anonymized supplemental material[6]), we performed retrieval experiments on the ProEdu database using the search interface. Each retrieved educational resource was matched to a gold standard reference set of expected items and precision, recall, and F1 were calculated to assess the efficacy of retrieval under multidisciplinary prompts.

[6] https://osf.io/wvkcm?view_only=06b075f8309840a8b7ce812df7ecb53a

5.2.4. Llama/Faiss Configuration and Experimental Procedure

We employ a RAG pipeline that integrates FAISS semantic retrieval (IndexFlatL2) with controlled generation by Llama using LangChain/ChatGroq. JSON records are transformed into documents, segmented (1000/200), and subsequently embedded utilizing all-MiniLM-L6-v2 with L2 normalization, thus employing cosine L2 measures. A persistent index retrieves the top k passages. The extracted passages are incorporated into the prompt, ensuring that Llama provides responses strictly based on the retrieved context. Detailed workflows are illustrated in Figure 3. Additional technical specifications concerning the orchestration and interaction mechanisms among the Llama model, FAISS, and ES are available in footnote at [7].

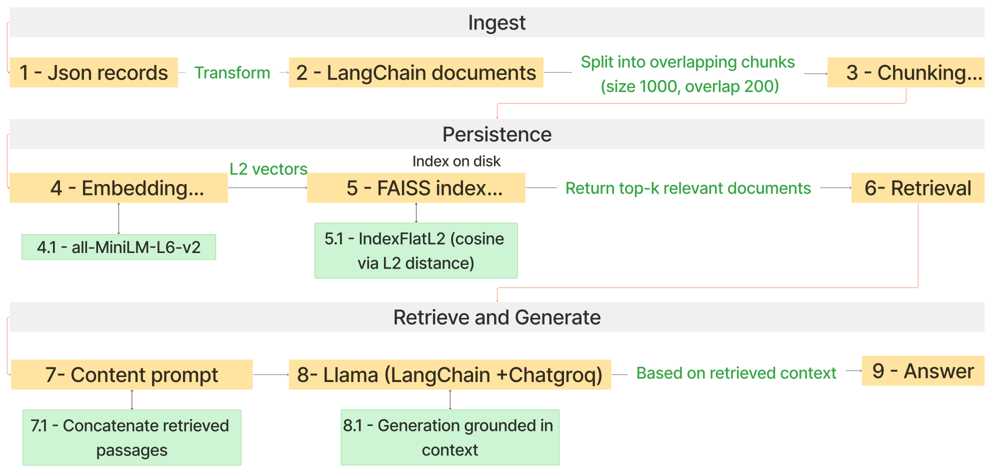

Figure 3. Diagram of llama/faiss configuration and experimental procedure

5.3. Evaluation

In the evaluation of the system, each prompt was paired with an established ground truth, comprising a collection of metadata pertinent to the applicable educational resources. Each metadata entry in this ground truth is organized with the following fields: resource ID, title, authorship, publication date, access URL, and associated keywords. The source code of the evaluation module[8] and the logging module[9] is provided in the footnotes to enable manual inspection of the experimental outcomes. Both the search engine responses and the computed evaluation metrics—precision, recall, and F1 score—were stored in a relational PostgreSQL database. The structure of the template file, which defines the set of 22 prompts and their associated educational resources, is shown in Listing 1. The evaluation was performed solely on the basis of the metadata available in the Proedu repository.

5.3.1. Validation

Two distinct strategies are employed based on the evaluation trajectory: in the instance of a user query, subjects are incorporated from ground-truth items whose subject fields

[7] https://osf.io/z573x/files/jmzec?view_only=01ad3cb5deff411b87f2a9cce5fe5cf5
[8] https://osf.io/e6wgd?view_only=06b075f8309840a8b7ce812df7ecb53a
[9] https://osf.io/x69wu?view_only=06b075f8309840a8b7ce812df7ecb53a

(ranging from subject1 to subject6) demonstrate similarity to the query. In the event of a prompt, the most analogous ground-truth prompt is selected, and the subjects of its items are aggregated. In both scenarios, subjects are normalized by the function clean_text. For each contextual document, methodically collect the normalized labels from the predefined fields, spanning subject1 to subject6.

The curated metadata of the repository is regarded as a highly reliable reference. The gold set was developed according to a written protocol [Li and Klabjan 2024], with the artifacts anonymized. To ensure internal validity and reproducibility, we meticulously document prompts, preprocessing, and hyperparameters (k, chunk size, embedding model). The intentional concentration on a single repository and language is designed to evaluate the proposal within a realistic framework, while the provision of transparent materials facilitates independent replication and further enhancements.

5.4. Eca Interface and Architecture

This section expounds on the modular architecture of the ECA and the data flow process, ranging from content acquisition to response formulation, that underpins its search and recommendation mechanism driven by generative artificial intelligence. As depicted in the flow diagram presented in Figure 4, the methodology applied in the development of the educational ECA encompasses seven primary stages executed sequentially: 1. Data organization, 2. User profile consolidation, 3. Processing (comprising text preprocessing and encoding), 4. Post-processing, 5. Document Filtering, 6. Model Output (including response generation), and 7. Feedback.

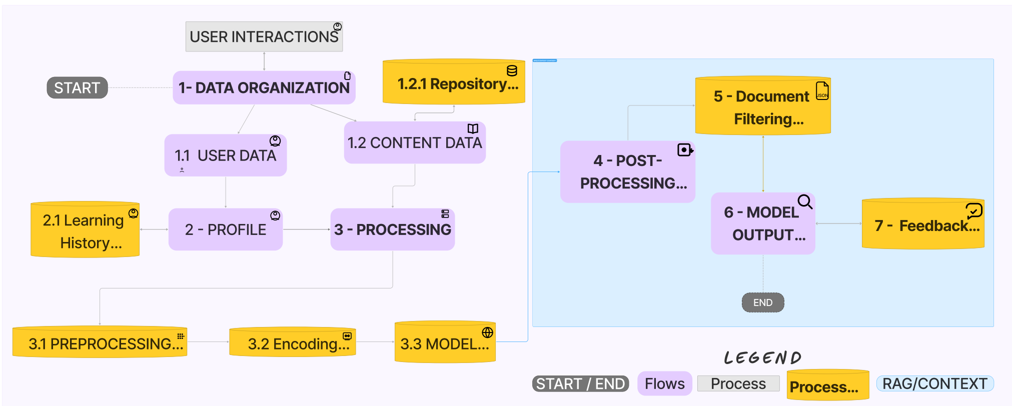

Figure 4. Flowchart of educational eca

The interface of the ECA system was meticulously designed with an emphasis on usability, embodying an intuitive layout. It is systematically organized into a side panel dedicated to session functionalities, including conversation history and authentication, as well as a primary content area. Within this configuration, tab navigation facilitates user interaction to alternate between various tools available, such as AI search (Llama + FAISS) and ES search. The display of search results is methodically structured, presenting essential metadata of each educational resource, such as title, date, access link, and subjects, thus simplifying the identification and retrieval of the intended material. For the presentation layer, a responsive web interface was developed utilizing the Streamlit framework [10], with

[10] https://docs.streamlit.io

the integration of the PyTorch library to coordinate data flow and model interactions [11].

Data organization is based on two interrelated sources. User data includes interaction logs, preferences, knowledge levels, browsing patterns, and evaluations. Content data is methodically cataloged in a repository designed to support retrieval-augmented generation, thereby enabling the storage of metadata and multimodal objects in PDF and TXT formats. User records are assimilated into a dynamic profile that encapsulates learning history, preferences, and behavioral patterns, establishing the essential foundation for the personalization of future stages.

The text-to-vector conversion process facilitates search operations that are cognizant of semantic meaning. This method begins with preprocessing, encompassing normalization and tokenization, proceeds to encoding, characterized by embeddings, and advances to model processing, wherein transformers, attention mechanisms, and prompt engineering enhance the refinement of semantic representations. Subsequently, post-processing endeavors to standardize representations, mitigate noise, and configure vectors for retrieval purposes. Retrieval of documents involves querying the RAG index and performing candidate filtration.

Afterward, a final post-processing stage renders responses in natural language, applying corrections and formatting for clarity and cohesion. The system outputs two durable artifacts: a knowledge base that powers ECA real-time interactions and a semantic search index exposed in the interface for source verification.

5.5. Reproducibility Guide

The source code of the application, along with the dataset utilized, tool validation files, ES index creation scripts, and algorithm implementation modules, is accessible at the electronic address indicated in the footnote [12]. The results of the metrics per prompt are accessible.[13] The inference of the Llama model is executed externally via the Groq API [14], with the droplet acting solely as the orchestrator, avoiding dedicated GPU use and ensuring predictable budgeting with future scalability margins.

5.6. Ethical And Compliance Statement

Generative use declaration: Generative AI tools were used for grammar and clarity checks. Tools: ChatGPT[15] and TexGPT. [16]

6. Summary of Results

The experimental results have been meticulously arranged into tables, with each of the 22 prompts assessed according to precision, recall, and F1 score metrics derived from the three evaluated search mechanisms. Comprehensive versions of these tables are accessible via the address specified in the footer. To facilitate interpretation, comparative graphs have also been generated to encapsulate the performance of the prompts across the three methodologies, offering an objective depiction of the observed patterns. The subsequent portion presents a critical analysis of these findings.

[11] https://pytorch.org
[12] https://osf.io/4rjn5?view_only=06b075f8309840a8b7ce812df7ecb53a
[13] https://osf.io/k82y4?view_only=06b075f8309840a8b7ce812df7ecb53a
[14] https://groq.com/
[15] https://chatgpt.com
[16] https://www.writefull.com

Table 4. Average performance on 22 prompts (mean ± SD; macro-average per prompt)

Algorithm	Precision	Recall	F_1
llama-3.3-70b-versatile + FAISS	0.83 ± 0.30	$\mathbf{0.78 \pm 0.23}$	$\mathbf{0.73 \pm 0.22}$
Elasticsearch	$\mathbf{0.86 \pm 0.35}$	0.66 ± 0.37	0.72 ± 0.35
ProEdu (lexical)	0.04 ± 0.12	0.03 ± 0.08	0.03 ± 0.07

6.1. Experimental Prompts

Table 3 provides a comprehensive overview of the distribution of 22 prompts across various subject areas, highlighting our methodical approach to crafting prompts. Each prompt has been meticulously designed to replicate authentic information retrieval queries within specified domains, encompassing a range from broad inquiries (such as retrieving all materials in Computer Science) to more narrowly focused queries with domain-specific constraints (such as restricting results to recent publications in molecular genetics). This extensive range and diversity in query construction are directly aligned with our objective: to rigorously evaluate the retrieval performance of the three selected algorithms (ES, ProEdu , and Llama) under diverse conditions.

By encompassing ten distinct domains of knowledge and incorporating both extensive searches and specific filtering criteria, the set of prompts ensures a rigorous evaluation of each algorithm's capacity to handle both broad and precision-specific information requirements. This comprehensive design of prompts not only spans a wide array of thematic content but also examines algorithms' accuracy in balancing recall with precision. These components are quantitatively evaluated using established metrics of assessment: precision, recall, and F1 score.

6.2. Discussions

As illustrated in Table 4, the llama-3.3-70b-versatile model in conjunction with FAISS achieved the highest recall rate and an elevated mean F1 score (0.73 ± 0.22). In contrast, the local ES baseline, while attaining the highest mean precision (0.86 ± 0.35), did so at the expense of a marked reduction in recall (0.66 ± 0.37). The solution currently implemented in ProEdu exhibited significant deficiencies in the retrieval of relevant resources.

The relatively large standard deviations for F1, precision and recall indicate considerable variability between the 22 prompts, which is likely due to the interdisciplinary nature of the queries. Ultimately, llama-3.3-70b-versatile + FAISS sustained high recall without unduly compromising precision, while the ES baseline emphasized precision by retrieving a smaller set of documents. To ensure replicability, the compilation script and data associated with the metrics a are provided in the footnote [17] [18].

In conclusion, the results of this study affirm the superior performance of the llama-3.3-70b-versatile + FAISS hybrid semantic approach in search tasks within educational repositories, offering an optimal balance between coverage and precision. Local ES is a feasible alternative when absolute precision is paramount, while the ProEdu solution still lags behind and requires substantial refinement to close the observed performance gap.

[17] https://osf.io/b84ds?view_only=06b075f8309840a8b7ce812df7ecb53a

[18] https://osf.io/cpefv?view_only=06b075f8309840a8b7ce812df7ecb53a

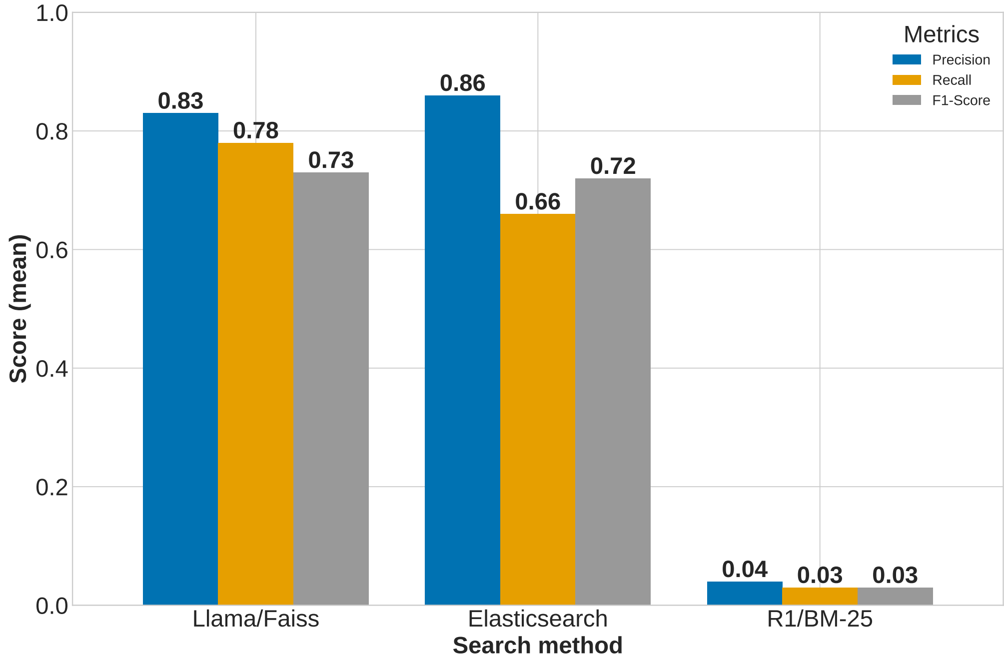

Figure 5. Mean of metrics by algorithm

The quality of the retrieval is a function of the interplay between three distinct dimensions, rather than a singular determinant. Metadata quality, which includes aspects such as coverage, accuracy, and normalization into controlled vocabularies (e.g., domain, grade level, keywords), facilitates precise filtering and scoring. The complexity of semantics, gaged by the number of concepts and relationships articulated in a query, elevates the requirement for reliable and structured signals beyond mere full text. Terminological alignment, characterized by the correspondence between the language of the user and the index vocabulary, dictates the efficacy of query-document matching.

The ES system exhibits a high degree of precision in the execution of concise factual search tasks. However, it displays a diminished capacity for recall when faced with queries that involve synonyms or paraphrased language. In contrast, Llama, when integrated with FAISS, maintains significant recall capacities. However, it is limited in addressing boundary query situations due to its reliance on generic embeddings. Lastly, the ProEdu methodology, inherently restricted by its search approach, is ineffective in scenarios that go beyond literal matching criteria.

6.3. Impact of F1-score

In the context of assessing the significance of the error, insufficient recall is considered more detrimental within educational environments. A system characterized by high precision but limited recall could potentially overlook vital resources.

In contrast, a slight surplus of irrelevant results, as indicated by high recall in conjunction with moderate precision, allows educators to exercise their pedagogical judgment in filtering content, particularly when the interface facilitates relevance-based sorting. Consequently, the superior F1 score obtained by the llama-3.3-70b-versatile combined with the FAISS-based retrieval pipeline demonstrates a more favorable balance between precision and recall. This enhanced trade-off provides more effective support for instructional content curation and reduces the manual effort required for content retrieval. The corresponding results are presented in Figure 5.

7. Contributions and Impact to IS area

This research involved a critical examination of three paradigms within the domain of information retrieval, specifically located in the context of an OER. It elucidated their technological underpinnings, evaluated their efficacy across interdisciplinary queries, and examined the pedagogical implications thereof. The semantic processing pipeline, which incorporates Llama and FAISS along with the advanced implementation of ES, demonstrated superior performance relative to the lexical mechanism of ProEdu .

This research paper introduces a significant advancement in the domain of Information Systems, particularly within the context of educational repositories, through the development of an architecture for semantic retrieval. This architecture integrates RAG with LLMs, specifically employing Llama and FAISS, and incorporates lexical signals using a conversational ECA interface.

In contrast to traditional lexical search engines such as ES, this method proficiently captures synonyms, paraphrases, and complex intents, thereby improving coverage and recall while maintaining an optimal equilibrium between precision and relevance, as evidenced by the F1 score.

This advancement significantly improves the current state of research in the discovery of OER and provides reproducible artifacts such as benchmarks and scripts, which serve to support both rigorous evaluation and extensive adoption. This study improves existing advances in semantic retrieval architectures, addresses pragmatic challenges within OER repositories, and addresses enduring research deficiencies in the domain of AI in education.

The practical implications are substantial: the system effectively addresses the persistent challenges faced by educators and learners in the identification of pertinent OER from vast and varied collections, which are frequently characterized by insufficient metadata and inconsistent terminologies. Improve search difficulties by producing conversational responses with verifiable sources, thus enhancing instructional curation and broadening opportunities for knowledge acquisition.

The results of this investigation demonstrate that the incorporation of generative models, exemplified by Llama, in conjunction with FAISS leads to notable progress in the field of educational information retrieval, especially in scenarios demanding a deep semantic understanding of the content. This methodological approach not only enhances the accuracy and relevance of search results but also guarantees the retrieval of a wider array of educational materials, thus improving accessibility to diverse learning resources.

The observed improvements in the F1 score across the evaluated systems indicate that an advanced contextually aware search mechanism, such as the Llama/FAISS pipeline, effectively mitigates the shortcomings inherent in conventional lexical-based methods. This is particularly significant within academic environments where both students and educators require dynamic, personalized content retrieval to meet varied learning needs. From a comprehensive educational and technical perspective, the findings underscore the transformative potential inherent in the integration of AI-driven semantic search engines into educational repositories.

In addition, the scalability of these systems ensures that institutions with limited resources can access advanced AI technologies. In conclusion, the incorporation of cutting-edge generative models within RAG pipelines represents a critical vector for innovation in open educational repositories, contingent on the presence of robust technological infrastructures, validated evaluation protocols, and an unwavering commitment to ethical and transparency standards. Future research efforts will involve rigorously structured investigations involving educators in the primary, secondary, professional, and technical domains to evaluate both the educational value and the operational efficacy of the ECA. Systematic questionnaires will be deployed to systematically document educators' interactions and experiences with the ECA, which compiles a wide array of open educational resources to support pedagogical practices and academic inquiries.

References

Boscarioli, C., de Araujo, R. M., and Maciel, R. S. P., editors (2017). *I GranDSI-BR: Grand Research Challenges in Information Systems in Brazil (2016–2026)*. Sociedade Brasileira de Computação (SBC), Brazil. E-book by SBC's Special Committee on Information Systems (CE-SI).

Bucchiarone, A., Bianculli, D., Zhang, Q., Pradella, M., Szilagyi, G., and Visaggio, M. (2024). Designing and generating lesson plans combining open educational content and generative ai. In *Companion Proc. 27th ACM/IEEE Int. Conf. on Model Driven Engineering Languages and Systems (MODELS Companion '24)*. ACM.

da Silva, F. L. and Cazella, S. C. (2020). Um modelo de recomendação de objetos de aprendizagem baseado em valores culturais dos estudantes. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 16–20, Porto Alegre, RS, Brasil. Sociedade Brasileira de Computação (SBC).

Dehbozorgi, N., Kunuku, M. T., and Pouriyeh, S. (2024). Personalized pedagogy through a llm–based recommender system. In *Artificial Intelligence in Education — Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks*, Communications in Computer and Information Science, pages 63–70. Springer Nature Switzerland.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvásy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. (2025). The faiss library. *arXiv preprint arXiv:2401.08281*.

Dutta, S., Beier, F., and Werth, D. (2025). AI-based personalized multilingual course recommender system using large language models. In *Proc. 17th Int. Conf. on Agents and Artificial Intelligence (ICAART 2025), Vol. 3*, pages 1069–1076. SCITEPRESS.

Face, H. (2021). all-minilm-l6-v2. Online repository. Available at: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed on: Oct. 7, 2025.

Gheewala, S., Xu, S., and Yeom, S. (2025). In-depth survey: Deep learning in recommender systems—exploring prediction and ranking models, datasets, feature analysis, and emerging trends. *Neural Computing and Applications*. Early access article, published 21 Mar 2025.

Golda, A., Mekonen, K., Pandey, A., Singh, A., Hassija, V., Chamola, V., and Sikdar, B. (2024). Privacy and security concerns in generative ai: A comprehensive survey. *IEEE Access*, 12:48126–48144.

Gormley, C. and Tong, Z. (2015). *Elasticsearch: The Definitive Guide*. O'Reilly Media, Sebastopol, CA.

Gregor, S. and Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2):337–355.

Kamphuis, C., de Vries, A. P., Boytsov, L., and Lin, J. (2020). Which BM25 do you mean? a large-scale reproducibility study of scoring variants. In *Advances in Information Retrieval (ECIR 2020)*, volume 12036 of *Lecture Notes in Computer Science*, pages 28–34. Springer.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Li, H. and Klabjan, D. (2024). Reverse prompt engineering. *arXiv preprint arXiv:2411.06729*. Available at: https://arxiv.org/abs/2411.06729.

Li, X., Jin, J., Zhou, Y., Zhang, Y., Zhang, P., Zhu, Y., and Dou, Z. (2025). From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62.

Li, X., Lipp, J., Shakir, A., Huang, R., and Li, J. (2024). Bmx: Entropy-weighted similarity and semantic-enhanced lexical search. *arXiv preprint arXiv:2408.06643*.

Mićunović, M., Rako, S., and Feldvari, K. (2023). Open educational resources (oers) at european higher education institutions in the field of library and information science during COVID-19 pandemic. *Publications*, 11(3):38.

Pesovski, I., Santos, R., Henriques, R., and Trajkovik, V. (2024). Generative ai for customizable learning experiences. *Sustainability*, 16(7):3034.

Sakai, T. (2021). *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. Springer. Open access.

Son, H. X., Nguyen, T. M., Vo, H. K., Dang, K. T., Gia, K. H., and Tran, N. B. (2024). Generative ai–driven digital assistance for e-learning: A novel paradigm for personalized recommendations. In *Artificial Intelligence with and for Learning Sciences. Past, Present, and Future Horizons (WAILS 2024)*, volume 14545 of *Lecture Notes in Computer Science*, pages 89–98. Springer.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Misra, I., Zhang, T., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. Meta AI Technical Report.

Zmaranda, D. R., Moisi, C. I., Győrödi, C. A., Ş. Győrödi, R., and Bandici, L. (2021). An analysis of the performance and configuration features of mysql document store and elasticsearch as an alternative backend in a data replication solution. *Applied Sciences*, 11(24).