Computers and Education: Artificial Intelligence 8 (2025) 100417

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Retrieval-augmented generation for educational application: A systematic survey

Zongxi Li [a,*], Zijian Wang [b], Weiming Wang [b], Kevin Hung [b], Haoran Xie [a], Fu Lee Wang [b]

[a] School of Data Science, Lingnan University, Tuen Mun, Hong Kong, China
[b] School of Science and Technology, Hong Kong Metropolitan University, Homantin, Hong Kong, China

ARTICLE INFO

Keywords:
Artificial intelligence in education
Retrieval-augmented generation (RAG)
Educational applications
Large language models (LLMs)

ABSTRACT

Advancements in large language models (LLMs) have transformed AI-driven education, enabling innovative applications across various learning and teaching domains. However, LLMs still face several challenges, including hallucination and static internal knowledge, which hinder their reliability in educational settings. Retrieval-Augmented Generation (RAG) enhances LLMs by retrieving relevant information from an external knowledge base and incorporating it into the LLM's generation process. This approach improves factual accuracy and enables dynamic knowledge updates, making LLMs particularly suitable for educational applications. In this paper, we comprehensively review existing research that integrates RAG into educational scenarios. We first clarify the definition and workflow of RAG, and following the indexing mechanism of RAG, we introduce different types of retrievers and generation optimization methods. As the main focus of this work, we explore the practical applications of RAG in education, covering interactive learning systems, generation and assessment of educational content, and large-scale deployment in educational ecosystems. Based on our comprehensive review, this paper discusses existing challenges and future directions, including mitigating hallucinations, ensuring the completeness and timeliness of retrieved knowledge, reducing computational costs, and enhancing multimodal support for RAG-based educational applications.

1. Introduction

The rapid advancement of large language models (LLMs), such as GPT-4, LLaMA and Gemini, has reshaped the landscape of artificial intelligence (AI) (Minaee et al., 2024). These models, trained on vast corpora of text data, exhibit remarkable capabilities in natural language understanding and generation. By leveraging deep neural networks and massive scale training, LLMs can generate coherent, contextually relevant, and human-like responses across a wide range of topics. Their general-purpose nature has led to widespread adoption in numerous fields, including healthcare, finance, scientific research, law and education, highlighting their adaptability across specific domains (Zhao et al., 2023; Chen et al., 2023). In the field of education, AI has been widely adopted (Chen, Xie, Zou, et al., 2020; K.F. Chen et al., 2024), and LLMs have been increasingly explored to enhance various learning and teaching processes. Researchers and educators have investigated their potential in both academic research and commercial tools (S. Wang et al., 2024). For instance, LLM-powered chatbots can serve as virtual tutors, providing students with real-time explanations and answering their questions (Prihar et al., 2023; Liang et al., 2024). Additionally, LLMs have been used in automated assessment, where they analyze student essays and provide feedback (Jeon & Lee, 2023). Their ability to process vast amounts of educational materials also makes them valuable for generating quizzes, summarizing course content, and assisting educators in lesson planning.

However, despite these advancements, current applications face critical challenges when deployed in real-world educational settings (Hwang et al., 2020; Stockwell, 2022). A major concern is the hallucination problem, where models generate factually incorrect or misleading information due to their probabilistic nature (Z. Li et al., 2025; Ji, Lee, et al., 2023; Ji, Liu, et al., 2023). In educational contexts, where accuracy and trustworthiness are essential, such issues can significantly impair learning outcomes. Another limitation is the static knowledge embedded within LLMs, which makes them unable to reflect the latest curricular updates or scientific advancements (Mallen et al., 2023; Meng et al., 2022; Zhang et al., 2024). Furthermore, LLMs often lack explainability and personalization, failing to meet diverse learner needs and

[*] Corresponding author.
E-mail address: zongxili@ln.edu.hk (Z. Li).

https://doi.org/10.1016/j.caeai.2025.100417
Received 28 March 2025; Received in revised form 2 May 2025; Accepted 6 May 2025
Available online 14 May 2025

leading to reduced trust in AI-supported learning systems (Zhao et al., 2024).

To address the limitations mentioned above, researchers have introduced Retrieval Augmented Generation (RAG), a hybrid framework that combines the generative power of LLM with an external retrieval mechanism (Lewis et al., 2020). Different from standard LLMs, which rely solely on their pre-trained knowledge, RAG first retrieves relevant documents from an external knowledge base before generating responses. This approach improves factual accuracy, knowledge freshness, and transparency, making it particularly suitable for educational applications where precise and verifiable information is crucial. By integrating retrieval-based knowledge augmentation, RAG systems can provide students with explanations, adapt to curriculum changes, and mitigate the risks of misinformation.

Recognizing the growing interest in RAG for education, this paper aims to provide a comprehensive review of its technical foundations, practical implementations, and potential impact on learning and teaching. To structure the discussion and highlight key research trends, we organize the survey around the following core research questions. Specifically, this survey aims to answer following research questions:

RQ1: How is RAG applied to enhance interactive and personalized learning in education?

RQ2: How does RAG support the development and assessment of educational content?

RQ3: How is RAG scaled and integrated within educational ecosystems?

1.1. Structure of the paper

The remainder of this paper is structured as follows. Section 2 provides a detailed overview of RAG, including its overall introduction, indexing mechanisms (Section 2.1), retrieval strategies (Section 2.2), and generation techniques (Section 2.3). This section establishes the technical foundation necessary for understanding RAG's role in educational applications. Section 3 describes the material and methods used to collect and screen the relevant literature, ensuring a systematic and comprehensive review. Section 4 explores the practical implementations of RAG in education, categorizing applications into three key domains: interactive learning systems (Section 4.1), educational content development and assessment (Section 4.2), and large-scale educational ecosystems (Section 4.3). Each subsection presents relevant studies and corresponds to a research question. Section 5 outlines the existing challenges and future directions of RAG-based educational systems, addressing issues such as hallucination, knowledge completeness, computational efficiency, and multimodal content integration. Finally, Section 6 concludes the paper by summarizing key findings, emphasizing the benefits of RAG in education, and highlighting areas for further research and development.

2. Retrieval-augmented generation

RAG is an advanced AI paradigm that integrates information retrieval (IR) with generative models to enhance the response accuracy and reliability (Lewis et al., 2020). Traditional IR systems efficiently locate relevant documents but cannot generate new content, while LLMs generate fluent text but rely solely on pre-trained knowledge. RAG bridges this gap through a two-step process: first retrieving pertinent documents from external databases, then leveraging them to guide text generation (Karpukhin et al., 2020). As illustrated in Fig. 1, the RAG workflow consists of three key stages: indexing, where documents are stored in a database; retrieval, where relevant documents are fetched based on user queries; and generation, where an LLM synthesizes responses using the input prompt combining the user query with relevant documents. This retrieval-augmented mechanism enables RAG to incorporate up-to-date knowledge, which in turn improves factual accuracy, contextual relevance, and adaptability. As a result, RAG is particularly effective for knowledge-intensive tasks that demand precise and reliable information (Izacard & Grave, 2021).

2.1. Indexing

Indexing aims to transform raw text into a structured and searchable format, enabling efficient and accurate retrieval of relevant information in the RAG framework. This process consists of several key stages, including data preprocessing, text chunking, vectorization, and index storage (Gao et al., 2023).

The first step in indexing is data preprocessing, which aims to extract clean and structured text from raw data sources. Educational materials such as textbooks (Gong et al., 2024), syllabus (Taneja et al., 2024), and exercise sheets (Neumann et al., 2025), these educational materials often exist in diverse formats including PDFs, PowerPoints, and Word documents (Chondamrongkul et al., 2025). To standardize these inputs, preprocessing involves converting them into plain text while eliminating noise, such as special characters, redundant whitespace, and non-informative metadata. This step ensures that subsequent indexing operations work on a consistent and high-quality textual representation.

Once the text is preprocessed, it is segmented into manageable units to facilitate effective retrieval, as language models have a limited context length (Das et al., 2025). This step involves splitting documents into smaller, self-contained segments, called chunks that can be processed independently. After chunking, each text segment is converted into a numerical representation through vectorization, enabling similarity-based retrieval. This process utilizes embedding models, such as BGE models (Zhang et al., 2020), ESE (X. Li et al., 2025), and OpenAI's text-embedding-ada-002,[1] to map textual content into a dense vector space. The selection of an appropriate embedding model depends on factors such as computational efficiency and domain-specific performance. Compared to traditional keyword-based methods, vector embeddings can capture the semantic relationships between words and concepts.

Finally, the processed text and its corresponding vector representations can be efficiently organized for retrieval. This is achieved through index storage, where each text chunk is assigned a unique identifier, such as a hash or document ID, and stored alongside its vector representation. To support fast and scalable retrieval, specialized vector databases, such as FAISS (Johnson et al., 2019), Weaviate,[2] and Pinecone,[3] are commonly employed. These databases enable approximate nearest neighbor searches, allowing for rapid retrieval of semantically similar information.

2.2. Retrieval

The retrieval phase is responsible for identifying and retrieving the most relevant text chunks from the indexed database to enrich the contextual knowledge of the language model. The quality and efficiency of retrieval play a crucial role in determining the accuracy and relevance of the generated response. Retrieval methods can generally be categorized into sparse retrieval, dense retrieval, and hybrid approaches.

2.2.1. Sparse retrieval

Sparse retrieval methods rely on traditional lexical matching techniques and sparse vector representations to retrieve relevant documents. One of the most widely used sparse retrieval methods is BM25, which employs TF-IDF weighting and probabilistic ranking functions to estimate the relevance between documents (Robertson et al., 2009). Due to its efficiency and interpretability, BM25 continues to be a prevalent

[1] https://platform.openai.com/docs/guides/embeddings.
[2] https://weaviate.io.
[3] https://www.pinecone.io.

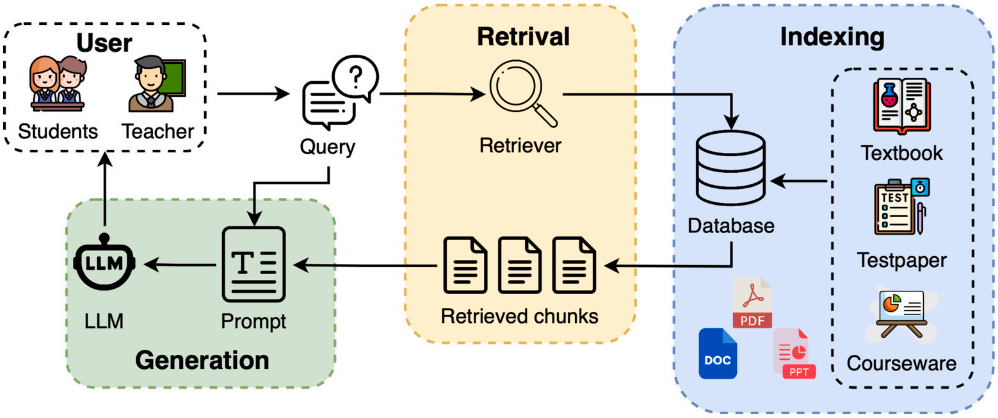

Fig. 1. The workflow of RAG.

method for RAG in various studies (Jiang et al., 2023; Ram et al., 2023a; Cheng et al., 2023). Beyond BM25-based methods, K-Nearest Neighbors (KNN) search has been explored for sparse retrieval. In contrast to BM25, which relies on exact lexical matches, KNN measures the similarity between query vectors and indexed document vectors based on distance metrics. Several studies have incorporated KNN-based retrieval methods to enhance performance in different retrieval scenarios (Alon et al., 2022; Borgeaud et al., 2022).

2.2.2. Dense retrieval

In contrast to sparse retrieval, dense retrieval methods leverage neural embedding models to transform both queries and documents into high-dimensional vector representations. A prominent example is Dense Passage Retrieval (DPR), which employs a dual-encoder architecture: one encoder processes queries, while another encodes documents (Karpukhin et al., 2020). Relevance is determined using inner product or cosine similarity, allowing for more nuanced semantic retrieval. Beyond DPR, other models have been developed for dense retrieval tasks. ColBERTv2 introduces late interaction mechanisms, enabling token-level similarity matching while maintaining computational efficiency (Santhanam et al., 2022). It has been utilized in various retrieval tasks, including applications described in Borgeaud et al. (2022) and Izacard et al. (2023). Contriever leverages contrastive learning to enhance representation robustness, particularly in zero-shot and few-shot retrieval settings (Izacard et al., 2022). It has been applied in studies such as Siriwardhana et al. (2023) and Sachan et al. (2021).

Despite its advantages, dense retrieval comes with notable computational challenges. Distinct from sparse retrieval methods, which rely on inverted indexes based on term frequencies, dense retrieval requires continuous embedding updates and high-dimensional similarity computations. These processes increase memory consumption and computational overhead, making dense retrieval more resource-intensive for large-scale applications.

2.3. Generation

The generation phase in RAG is responsible for synthesizing responses by integrating retrieved information with LLMs. The primary goal of this phase is to ensure that the generated output is accurate, coherent, and contextually relevant. This section discusses key aspects of the generation process, including retrieved information processing and fusion techniques, as well as optimization strategies for LLMs.

2.3.1. Processing and fusion of retrieved information

The effectiveness of RAG relies on how well the retrieved information is processed and integrated into the generation process. A straightforward approach is concatenation-based generation, where retrieved text chunks are directly appended to the user query before being fed to the LLM (Lewis et al., 2020). While simple, this method can lead to redundancy or exceed the model's context length when handling a large number of retrieved documents. To improve the efficiency of information integration, some approaches apply knowledge aggregation, which involves relevance filtering (Asai et al., 2023), information compression (Xu et al., 2024), and re-ranking (Glass et al., 2022) before passing the retrieved text to the LLM. Another technique, Fusion-in-Decoder (FiD), processes retrieved documents separately, allowing the model to weigh their contributions dynamically during the response generation process. Contrary to simple concatenation, FiD enables the model to evaluate different evidence sources in parallel and synthesize a more coherent output (Izacard & Grave, 2021). Additionally, adaptive fusion generation methods leverage attention mechanisms or re-ranking strategies to dynamically control the influence of retrieved knowledge during decoding. These techniques aim to mitigate the risks of hallucination by ensuring the model prioritizes more relevant and reliable external information (Guu et al., 2020).

2.3.2. Optimization strategies for LLMs

To enhance the generation quality in RAG, various optimizations can be applied to LLMs. One fundamental strategy is prompt engineering, which designs structured prompts to guide the model's use of retrieved information. For instance, In-Context RALM maintains the model parameters in a frozen state while integrating retrieved information through prompt fusion techniques to enhance the generation process (Ram et al., 2023b). Another approach is fine-tuning, where LLMs are trained on domain-specific datasets to better grasp specialized terminology and ensure contextual relevance (Cheng et al., 2023). Fine-tuning on retrieval-augmented datasets can also enhance the model's ability to integrate retrieved knowledge effectively. Lastly, hybrid generation strategies combine rule-based methods with generative approaches to enhance output reliability. For example, some systems generate multiple candidate responses and use a re-ranking mechanism to select the most accurate answer (Karpukhin et al., 2020). In structured tasks, integrating knowledge graphs or database queries alongside LLM-generated text can further improve factual consistency.

3. Material and methods

3.1. Article collection phase

To systematically collect relevant studies, we utilize the Web of Science (WoS) database[4] based on several considerations. First, the WoS is a widely used database for bibliometric and systematic literature reviews, valued for its comprehensive indexing and citation tracking

[4] https://webofknowledge.com.

(Roemer & Borchardt, 2015). Second, WoS covers both computer science and education more broadly than specialized databases, making it well-suited for identifying interdisciplinary research at their intersection. Third, WoS has been extensively used in prior studies focused on the overlap between computer science and education (Chen, Xie, Zou, et al., 2020; Chen, Xie, & Hwang, 2020; Ng et al., 2021). Additionally, to minimize the risk of overlooking relevant studies, particularly recent conference papers or preprints that may not yet be indexed by WoS. We also conducted a complementary backward citation search using Google Scholar.

We formulate a search query that incorporates key terms that capture the concept of RAG, such as "retrieval augmented generation", "retrieval augmented generation", "retrieval based generation", and "retrieval-augmented models". These terms were combined with keywords relevant to the educational domain, including "education", "teaching", "student", and "classroom". The keywords "LLM" and "training" were intentionally excluded from the query to maintain retrieval precision. Since RAG inherently relies on LLMs as a core component for generation, explicitly including the keyword "LLM" in the search query would be redundant. Additionally, the term "training" is commonly associated with model training in computer science, which could introduce irrelevant studies unrelated to our focus. The search was conducted using the Topic (TS) field, ensuring that the retrieved articles contained these terms in their titles, abstracts, or keywords. The search query used was as follows:

TS = (("retrieval augmented generation" OR "retrieval-augmented generation" OR "retrieval-based generation" OR "retrieval-augmented models") AND ("education" OR "teaching" OR "student" OR "classroom"))

The search was carried out on 6 March 2025 and the publication period was restricted to 2020-2025, since RAG was first introduced as a concept in 2020 (Lewis et al., 2020).

3.2. Article screening phase

To further ensure relevance at the intersection of RAG and education, we refined our selection by removing studies that do not fall within the research domains of "Computer Science", "Education & Educational Research", and "Engineering". We then applied exclusion criteria to improve the relevance and quality of the included studies. Studies were excluded if they (1) only mentioned RAG in discussions of future work without any empirical implementation; (2) primarily focused on non-educational domains without direct application to educational contexts; or (3) were review articles rather than empirical or system development studies. Consequently, 45 articles were directly cited in our study.

Additionally, to enhance the comprehensiveness of our literature review, we conducted a backward citation search by analyzing the related work sections of the retrieved articles. This led to the inclusion of 6 additional relevant studies identified through Google Scholar, bringing the total number of cited articles to 51. These 6 studies were not retrieved in the initial WoS search primarily because they were not indexed in WoS, such as preprint servers or conference proceedings that do not meet WoS inclusion criteria.

4. RAG application in education

In this section, we explore the practical applications of RAG in the field of education. To systematically derive the summary of RAG applications in education, we followed a structured analysis aspects, as presented in Table 1. First, we extracted application-related information from the retrieved articles' titles, abstracts, methodology sections, and experimental evaluations. We focused specifically on identifying (1) the educational scenarios targeted, (2) the objectives of RAG integration, (3) the indexing and retrieval methods used, and (4) the language models adopted for generation.

Then, as illustrated in Fig. 2, we manually coded each article based on its primary application scenario into three major categories: (a) Interactive Learning Systems, (b) Educational Content Development and Assessment, and (c) Large-Scale Educational Ecosystem Deployments. In cases where an article covered multiple scenarios, we assigned it to the category corresponding to its most emphasized application. Within each main category, we further subdivided studies according to more specific application scenarios.

Additionally, to align with the research objectives of this survey, we mapped each major application category to a corresponding RQ to guide the analysis. Specifically, Section 4.1 addresses RQ1, exploring the various types of RAG-powered interactive learning systems and their design strategies. Section 4.2 responds to RQ2, focusing on how RAG enhances automated educational content generation and intelligent assessment mechanisms. Section 4.3 examines RQ3, investigating broader institutional-level implementations and technical advancements that enable large-scale deployment.

4.1. RAG-powered interactive learning systems (RQ1)

RAG-Powered Interactive Learning Systems use RAG to enable dynamic, personalized, and context-aware interactions in education. These systems are characterized by their ability to facilitate conversational engagement, retrieve domain-specific knowledge, and adapt to diverse educational scenarios.

For RQ1, this survey identifies four ways in which RAG is applied to enhance interactive and personalized learning in education: the Educational Q&A System, the Educational Chatbot, the AI-driven Tutoring System, and the Adaptive Learning Paths. These trends highlight the diverse ways in which RAG has been leveraged to create dynamic, context-aware, and personalized educational interactions.

4.1.1. Educational Q&A systems

Educational Q&A System aims to provide precise answers to students' course-related questions.

Subject-specific Q&A systems. In anatomy courses, Anatbuddy integrates a curated medical knowledge base to provide precise anatomical explanations (Arun et al., 2024). By grounding its responses in domain-specific resources, it outperforms general-purpose models like ChatGPT in accuracy and contextual relevance. Beyond medical applications, RAG has also been leveraged in supply chain management (SCM) courses. Researchers combined Google Vertex AI Search with the SCOR model to structure SCM-related queries, retrieving course materials and generating responses aligned with established SCM frameworks (Ehrenthal et al., 2024). This approach enhances educational effectiveness by ensuring that AI-generated answers adhere to domain-specific guidelines. Mathematics education has also benefited from RAG-powered Q&A systems. One study applied RAG to algebra and geometry, retrieving relevant textbook excerpts before generating responses (Henkel et al., 2024). The study further explored prompt engineering strategies, highlighting a trade-off between fully grounded explanations and flexible, context-aware responses preferred by human users.

MOOC course Q&A systems. In the context of Massive Open Online Courses (MOOCs), one study compared the performance of standard GPT-4, RAG-enhanced GPT-4, and human students in assessment tasks (Miladi et al., 2024a). The results showed that RAG-enhanced models performed better, particularly in fill-in-the-blank exercises. This study also introduced Persona Prompting, a framework that refines AI responses through role-specific instructions, ensuring better alignment with course objectives. Another MOOC-focused study developed a conversational agent that retrieves and structures the course materials in real time (Miladi et al., 2025). By reducing hallucinations and improving response reliability, this system enhanced students' knowledge retention compared to those using standard GPT-4.

Table 1
A Summary of RAG Application in Education.

Application	Usage Scenario	Objective	Indexing	Retrieval	Generation
AITeach (Chondamrongkul et al., 2025)	Educational Software System	Enhance content generation accuracy and applicability using study materials.	OpenAI-text-embedding-3-small HF Instructor model	Semantic Similarity Retrieval	Gemini-1.5-Pro-002 Gemini-1.0-Pro-002 GPT-3.5-Turbo GPT-4-Turbo
MoodleBot (Neumann et al., 2025)	Educational Chatbot	Improve the accuracy and credibility of the generated content using course materials.	OpenAI-text-embedding-ada-002	Semantic Similarity Retrieval	GPT-4
DSRAG (Fung et al., 2024)	Educational Feedback Generation	Generate personalized feedback using student records and assignments.	OpenAI-text-embedding model	Semantic Similarity Retrieval	GPT-4
Zheng et al. (2024)	Customized Lesson Plan Generation	Enhance content generation accuracy and applicability using lesson plans.	-	Query-Based Database Retrieval	GPT-4
Liu et al. (2024c)	Intelligent Teaching Assistant	Enhance content relevance and context understanding using CS1 coursewar.	-	Multi-layered Nested Retrieval	Mistral-7B
ChatEd (K. Wang et al., 2024)	Course-related Question Answering	Enhance content relevance and context understanding using teacher-provided course materials	-	Facebook AI Similarity Search library	GPT-3.5-Turbo
Anatbuddy (Arun et al., 2024)	Educational Chatbot	Improve answer quality using medical resources.	-	Semantic Similarity Retrieval	GPT-3.5
Miladi et al. (2024a)	MOOC in Online Education	Enhance the accuracy of generated content using AI domain-specific texts.	OpenAI-text-embedding-ada-002	Semantic Similarity Retrieval	GPT-4
E-OED (Bui et al., 2024)	Educational Q&A System	Provide context to improve Q&A accuracy using education policy and intention entities.	Sentence-BERT	Semantic Similarity Retrieval	URA
Lee (2024)	Statistics Education Tutoring	Reduce hallucinations using embedded course notes text blocks.	OpenAI-text-embedding-ada-003	Semantic Similarity Retrieval	GPT-3.5
EduChat (Dan et al., 2023)	Open Question Answering	Improve timeliness and accuracy using latest internet content.	-	Online Retrieval with Self-Checking	LLaMA-13B-Base
Ko et al. (2024)	Python Programming Skill Acquisition	Reduce cognitive load using Python programming materials	LLM-RAG-Gradio platform	Vector Similarity Retrieval	GPT
CollaBot (Hu et al., 2025)	Online Collaborative Writing	Generate quality feedback using domain knowledge and scaffolding strategies.	Sentence-BERT	Online Retrieval	Spark
Hennekeuser et al. (2024)	Higher Education Lecturing Assistance	Provide personalized responses using lecture materials.	OpenAI-text-embedding-ada-002	Semantic Similarity Retrieval	GPT-4
Miladi et al. (2025)	MOOC Learner Support	Improve accuracy and contextual relevance using explanatory texts and video transcripts.	OpenAI-text-embedding-ada-002 model	Semantic Similarity Retrieval	GPT-4
Jauhiainen and Guerra (2024)	Student Response Assessment	Provid context for student responses evaluation using student learning materials.	LangChain OpenAI Embedding	Semantic Similarity Retrieval	GPT-3.5 GPT-4 Claude-3 Mistral-Large
ALINet (Zeghouani et al., 2024)	Automated Question Generation	Improve question generation using reading material	-	Semantic Similarity Retrieval	BART-Base
OwlMentor (Thüs et al., 2024)	Scientific Text Comprehension	Reduce hallucinations and provide relevant responses using scientific text embeddings	OpenAI-text-embedding-ada-002 model	Logical Routed Retrieval	GPT-3.5-Turbo
Z. Chen et al. (2024)	University Admission Consulting	Improve response accuracy using college admissions Q&A pairs	Sentence-BERT	Semantic Similarity Reetrieval	Text-davinci-003 GPT-3.5-Turbo
Dakshit (2024)	Virtual Teaching Assistan	Improve response accuracy and reliability using slides and materials.	-	-	Gemini
SKYRAG (Soekamto et al., 2025)	Personalized Learning Path Generation	Enhance retrieval precision and learning path generation using MOOC courses.	Sentence Transformers all-mpnet-v2 variant	Semantic Similarity Retrieval and Keyword-driven filtering	LLaMA3.1-70B
HICON (Singla et al., 2024)	Higher Education Counseling Bot	Reduce hallucinations and improve generated content accuracy.	-	Semantic Similarity Retrieval	LLaMA2
HiTA (C. Liu et al., 2024)	Course assistance	Ensure generated content aligns with course materials.	-	Semantic Similarity Retrieval	GPT-3.5-Turbo GPT-4
KG-RAG (Dong, 2023)	AI Tutor	Improve the accuracy and consistency of AI teaching using course materials.	OpenAI-text-embedding-v2	Semantic Similarity Retrieval	DeepSeek-V3
Han et al. (2024)	Tutor Social-emotional learning Competency Assessment	Improve evaluation accuracy using tutoring transcriptions.	-	Semantic Similarity Retrieval	GPT-3.5 GPT-4

(*continued on next page*)

Table 1 (*continued*)

Application	Usage Scenario	Objective	Indexing	Retrieval	Generation
GenSCM (Ehrenthal et al., 2024)	Supply Chain Management Education	Enhance knowledge generation using supply chain PDF documents.	-	Google Vertex AI Search	PaLM 2 for Text
Aboukacem et al. (2024)	Student Dropout Prediction	Enhancing contextual understanding for student dropout prediction using student data.	-	Semantic Similarity Retrieval	Zephyr-7b-beta Mistral-7b
CVTutor (Teng et al., 2024)	Flipped Classroom	Enhance LLM knowledge using course material.	OpenAI-text-embedding-3-small	Semantic Similarity Retrieval	GPT-4-Turbo
Jill Watson (Taneja et al., 2024)	Virtual Teaching Assistant in Online Classrooms	Improve LLM reliability using course material.	-	Semantic Similarity Retrieval	GPT-3.5
LAMB (Alier et al., 2025)	AI Learning Assistants	Improve the accuracy and contextual relevance using knowledge blocks in semantic storage.	OpenAI-text-embedding-3-large	Semantic Similarity Retrieval	LLaMA2-13B LLaMA3-7B Mixtral-7B
Miladi et al. (2024b)	MOOCs Online Education	Improve response accuracy using multimodal educational content.	OpenAI-text-embedding-ada-002	Semantic Similarity Retrieval	GPT-4 GPT-3.5
Jacobs and Jaschke (2024)	Programming education Feedback System	Reduce hallucinations and enhance feedback specificity using course material.	OpenAI-text-embedding-ada-002	Semantic Similarity Retrieval	GPT-4
MCQGen (Hang et al., 2024)	Personalized MCQ Generation	Improve quality and relevance of LLM-generated MCQs leveraging MCQs created by students and educators.	-	Semantic Similarity Retrieval	GPT-4
Lane (2025)	Geographical Education Support	Provide precise feedback and correct content.	-	-	-
Dehbozorgi et al. (2024)	Educational Recommender System	Improve contextual relevance in educational recommendations using pedagogical design patterns.	Transformer-based embedding models	Semantic Similarity Retrieval	LLaMA2-13B
RAMO (Rao & Lin, 2024)	Education Recommender System	Address cold-start issues and enhance recommendation relevance using course information.	OpenAI-text-embedding-ada-002	Semantic Similarity Retrieval	GPT-3.5-Turbo
Henkel et al. (2024)	Math Q&A System	Improve correctness and reliance on textbooks in math Q&A using textbook sources.	OpenAI-text-embedding-ada-002	Semantic Similarity Retrieval	GPT-3.5-Turbo-0613
Soliman et al. (2024)	Personalized Learning Support	Generate context-aware answers using learning materials.	OpenAI-text-embedding-3-small	Semantic Similarity Retrieval	GPT-3.5-Turbo
Capari et al. (2024)	Concept Definition Generation	Generate high-quality scientific definitions using books and review articles.	Msmarco-Distilbert-Base-tas-b	Semantic Similarity Retrieval	GPT-3.5
SiameseCamemBERT-Bio (El Malhi et al., 2024)	Medical Knowledge Retrieval	Enhancing semantic search and improving Q&A accuracy using anatomy books.	CamemBERT-Bio	Semantic Similarity Retrieval	CamemBERT-Bio
Barnett et al. (2024)	Research, Education, Biomedical	Reduce hallucinations using research papers and educational content.	-	Semantic Similarity Retrieval	GPT-4
SteLLA (Qiu et al., 2024)	Short-Answer Grading	Extract structured data from instructor-provided reference answers.	-	-	GPT-4-Turbo-Preview
Li et al. (2024)	Curriculum-Based Teaching	Improve matching of diagnostic tests with goals using curriculum data.	OpenAI-text-embedding-ada-002	Semantic Similarity Retrieval	GPT-3.5-Turbo-0613
Fernandez et al. (2024)	Course logistics-related question answering	Improve course-related Q&A accuracy using course syllabi.	-	BM25-Based Retrieval	LLaMA2-7B LLaMA2-13B LLaMA2-70B GPT-4
R. Liu et al. (2024)	AI-Enhanced CS50 Course Discussions	Reduce hallucinations and improve response accuracy using lecture captions.	OpenAI-text-embedding-ada-002	Semantic Similarity Retrieval	GPT-4
Gong et al. (2024)	Medical Question Answering	Retrieve latest research using medical textbooks and research papers.	-	BM25-Based Retrieval	GPT-4o
Guo and Li (2024)	AI Writing Assistance	Provide customized support.	-	-	-
Izquierdo-Domenech et al. (2024)	Information retrieval and Q	Combine LLMs and context-aware retrieval using teaching content	-	Keyword-Based Retrieval	GPT-3.5 Text-davinci-003
VizChat (Yan et al., 2024)	Learning Analytics Dashboards (LADs) Enhancement	Improve LLM-generated explanations and reduce hallucinations using instructional documentation	OpenAI-text-embedding-ada-002	Semantic Similarity Retrieval	GPT-4V

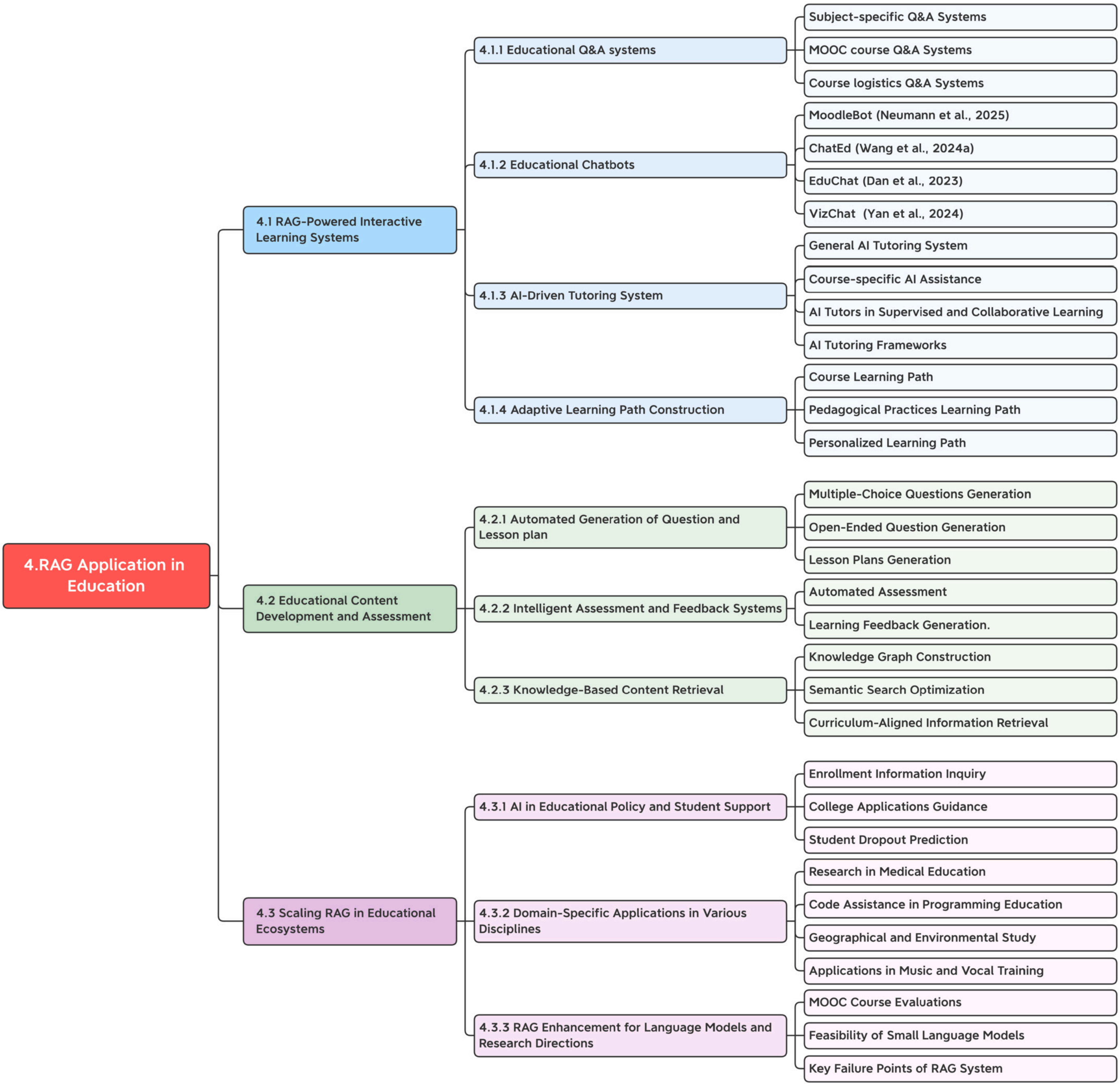
4.RAG Application in Education
4.1 RAG-Powered Interactive Learning Systems
4.1.1 Educational Q&A systems
Subject-specific Q&A Systems
MOOC course Q&A Systems
Course logistics Q&A Systems
4.1.2 Educational Chatbots
MoodleBot (Neumann et al., 2025)
ChatEd (Wang et al., 2024a)
EduChat (Dan et al., 2023)
VizChat (Yan et al., 2024)
4.1.3 AI-Driven Tutoring System
General AI Tutoring System
Course-specific AI Assistance
AI Tutors in Supervised and Collaborative Learning
AI Tutoring Frameworks
4.1.4 Adaptive Learning Path Construction
Course Learning Path
Pedagogical Practices Learning Path
Personalized Learning Path
4.2 Educational Content Development and Assessment
4.2.1 Automated Generation of Question and Lesson plan
Multiple-Choice Questions Generation
Open-Ended Question Generation
Lesson Plans Generation
4.2.2 Intelligent Assessment and Feedback Systems
Automated Assessment
Learning Feedback Generation.
4.2.3 Knowledge-Based Content Retrieval
Knowledge Graph Construction
Semantic Search Optimization
Curriculum-Aligned Information Retrieval
4.3 Scaling RAG in Educational Ecosystems
4.3.1 AI in Educational Policy and Student Support
Enrollment Information Inquiry
College Applications Guidance
Student Dropout Prediction
4.3.2 Domain-Specific Applications in Various Disciplines
Research in Medical Education
Code Assistance in Programming Education
Geographical and Environmental Study
Applications in Music and Vocal Training
4.3.3 RAG Enhancement for Language Models and Research Directions
MOOC Course Evaluations
Feasibility of Small Language Models
Key Failure Points of RAG System

Fig. 2. A Summary of RAG Application in Education.

Course logistics Q&A systems. Apart from answering questions about course content, the Q&A system also plays a role in course logistics. SyllabusQA introduces a publicly available dataset designed for course logistics question answering, such as course schedules and grading policies (Fernandez et al., 2024). In contrast with content-oriented Q&A systems, this study evaluates different retrieval strategies, including BM25 and fine-tuned LLaMA models, offering insights into optimizing structured academic document retrieval. In addition, RAG has been explored as a tool for AI-driven lecture support. One study embedded course materials into a Milvus vector database, enabling an AI system to assist university instructors by retrieving relevant teaching resources in response to faculty queries (Hennekeuser et al., 2024). The study emphasizes the importance of transparency and citation mechanisms to build trust in AI-generated content, particularly in academic settings.

4.1.2. *Educational chatbots*

Educational chatbots differ from Q&A systems by maintaining continuous, coherent conversations to support learning (Fan et al., 2024). One of the most comprehensive integrations of RAG is MoodleBot, an AI-powered chatbot embedded within a Learning Management System (LMS) (Neumann et al., 2025). By combining GPT-4 with Weaviate as a vector database, MoodleBot retrieves course-specific materials such as lecture notes and assignments to support self-regulated learning and help-seeking behaviors. Taking a hybrid approach, ChatEd integrates ChatGPT with traditional IR techniques (K. Wang et al., 2024). By employing PostgreSQL and Faiss for vector search, ChatEd ensures that retrieved content aligns with course materials, improving response accuracy and student satisfaction when compared to standard LLM-based chatbots. Beyond content retrieval, some chatbots incorporate pedagogical strategies to enhance learning outcomes. EduChat, designed for higher education, integrates psychological and pedagogical theories into its framework (Dan et al., 2023). It employs an online knowledge retrieval mechanism and adopts Socratic dialog techniques, fostering critical thinking and deeper engagement with learning materials. Expanding the scope of RAG applications, VizChat enhances the interpretability of Learning Analytics Dashboards (LADs) by leveraging multimodal generative AI (Yan et al., 2024). Dissimilar to conventional educational chatbots, VizChat provides contextualized explanations of visualized learning data, improving students' understanding of their progress. Additionally, by incorporating memory mechanisms, it enables the system to use prior interactions to inform and shape future responses, allowing for personalized and dynamic explanations.

4.1.3. AI-driven tutoring systems

The AI-Driven tutoring system is designed to provide personalized learning guidance and feedback to help users understand and master knowledge. Research on RAG-based AI tutoring systems has explored various approaches to improving student learning outcomes.

General AI tutoring system. One line of work focuses on AI tutors that integrate structured knowledge to enhance conceptual coherence. For instance, the KG-RAG framework combines knowledge graphs with RAG, enabling AI tutors to retrieve and generate educational content while maintaining logical consistency (Dong, 2023). By aligning AI-generated explanations with domain knowledge, this approach improves content accuracy and supports deeper learning. Another line of research examines the role of RAG-powered AI tutors in assisting students with academic research. OwlMentor, for example, helps university students navigate scientific literature by employing logical routed retrieval and the Annoy vector database to retrieve relevant academic content (Thüs et al., 2024). This system enhances student engagement with research papers, and its adoption is evaluated using the Technology Acceptance Model (TAM) to assess its impact on learning outcomes. A similar study investigates faculty perspectives on RAG-based AI tutors, highlighting the need for broader knowledge integration beyond course materials and improved handling of mathematical content (Dakshit, 2024). Beyond structured AI tutoring, some studies explore student-driven AI tutoring that encourages self-regulated learning. One study on English as a Foreign Language (EFL) learning examines how students create and customize their own RAG-powered chatbots for writing assistance (Guo & Li, 2024). Using the Poe platform, students upload personalized knowledge bases, such as writing guidelines and example essays, to refine chatbot-generated feedback.

Course-specific AI assistance. Several AI tutors have been developed to provide direct academic assistance to students by retrieving course-related materials. One example is Jill Watson, a virtual teaching assistant that answers student queries using DPR to extract relevant information from structured sources such as syllabi and lecture notes (Taneja et al., 2024). Its modular design allows for seamless integration via APIs, while safeguards like the OpenAI Moderation API and textual entailment checks enhance content reliability. Another course-specific AI tutor, CS50, is designed for Harvard's introductory computer science course (R. Liu et al., 2024). It integrates RAG with GPT-4 to provide programming guidance without directly offering solutions to assignments. Instead, it helps students refine their coding style, debug errors, and understand concepts, reinforcing problem-solving skills. In Python programming education, a study integrates RAG with LLMs to create an interactive learning platform, reducing cognitive load and improving programming proficiency (Ko et al., 2024). This system retrieves structured programming resources and integrates with the LMS to provide contextualized feedback. Similarly, an AI tutor designed for computer statistics education leverages ChatGPT, LlamaIndex, and Streamlit to assist students in understanding statistical concepts such as t-tests and ANOVA (Lee, 2024). By applying structured retrieval, it ensures responses remain grounded in course materials and minimizes hallucinations.

AI Tutors in Supervised and Collaborative Learning. AI tutors based on RAG have also been applied in interactive and collaborative learning environments. One approach integrates AI tutoring with teacher supervision, as demonstrated by HiTA, an AI-assisted learning platform designed to keep educators at the center of instruction (C. Liu et al., 2024). Teachers can oversee AI-generated responses, customize teaching materials, and track student interactions. HiTA offers multiple interaction modes, including standard Q&A, assignment assistance with hint-based guidance, and practice exercises aimed at promoting active learning while mitigating over-reliance on AI-generated content. In collaborative learning settings, AI tutors also facilitate students' peer interaction and feedback. CollaBot, an AI-powered conversational agent, supports students in online collaborative writing (OCW) by providing adaptive, multi-tiered scaffolding (Hu et al., 2025). Its feedback framework encompasses cognitive, meta-cognitive, and social dimensions, helping students organize, refine, and improve their writing, fostering deeper engagement with the content and collaborative process. Beyond structured collaboration, AI tutors also enhance self-directed learning in flipped classrooms. CVTutor, for instance, leverages RAG and multi-agent collaboration to assist students before and after in-person instruction (Teng et al., 2024). The system addresses a range of tasks, from answering course-related queries to grading assignments and ensuring ethical compliance in AI-generated content. Its integration with LMS streamlines the learning process, and empirical studies indicate its potential to improve both student performance and engagement.

AI Tutoring Frameworks. Scalability remains a critical factor in AI tutoring adoption, particularly for institutions with large student populations. Learning Assistant Manager and Builder (LAMB) is one such framework that enables institutions to develop AI learning assistants capable of seamless integration with LMS platforms (Alier et al., 2025). By leveraging RAG, LAMB ensures that AI-generated responses are grounded in authoritative sources, including lecture transcripts, PDFs, and academic papers. Additionally, the framework supports prompt engineering and plugin-based customization, allowing educators to tailor AI interactions to specific course requirements. Similarly, the BiWi AI Tutor is designed to provide scalable mentoring support in higher education (Soliman et al., 2024). It utilizes LangChain and LlamaIndex to facilitate structured retrieval from multiple educational sources, including lecture slides and seminar transcripts. The system dynamically selects the most relevant materials based on student queries, though the study notes challenges in optimizing retrieval accuracy and maintaining up-to-date knowledge bases.

4.1.4. Adaptive learning path construction

Adaptive learning path construction focuses on designing a personalized sequence of learning activities and resources, emphasizing the relationships between courses rather than the content within individual courses. One such approach is Retrieval-Augmented Generation for MOOCs (RAMO), which integrates a retrieval component with an LLM to enhance course recommendations, particularly addressing the "cold start" problem for new users who lack historical data (Rao & Lin, 2024). By utilizing OpenAI's text-embedding-ada-002 for vector-based retrieval and GPT-3.5-Turbo for content generation, RAMO is designed to provide personalized course suggestions without relying on prior user interactions. However, the study does not incorporate domain-specific educational models, which could further optimize recommendations for learning contexts. Additionally, the reliability of course recommendations is constrained by the dataset, which is primarily sourced from the Coursera Courses Dataset 2021, limiting its applicability to broader educational domains. Then, the Separated Keyword Retrieval Augmentation Generation (SKYRAG) introduces a keyword-driven retrieval mechanism, refining search precision by extracting and structuring user queries before retrieving relevant learning resources (Soekamto et al., 2025). Unlike RAMO, which primarily relies on vector-based retrieval, SKYRAG employs a hybrid approach that combines keyword extraction with semantic retrieval using Sentence Transformers. This method enhances the alignment between user queries and retrieved course materials, improving the coherence of generated learning pathways. Additionally, SKYRAG incorporates a structured learning path generation mechanism, considering prerequisite knowledge and thematic continuity, which is absent in RAMO. Another study focuses on leveraging RAG for personalized pedagogy by constructing a structured knowledge base of pedagogical design patterns (PDPs) (Dehbozorgi et al., 2024). Dissimilar to RAMO and SKYRAG, which primarily target student course recommendations, this research is designed to assist educators in selecting instructional strategies. The system employs a similarity search mechanism to retrieve the most relevant PDPs based on user queries and integrates them into teaching recommendations using Llama2-13B-Chat-H. A notable contribution of this study is its evaluation framework, which applies the Giskard framework to assess model accuracy, bias, and hallucination issues. Further advancing the concept of structured learning pathways, AITeach leverages a Chain-of-Reasoning-and-Action framework to deliver highly personalized educational experiences, meticulously tailored to the unique needs of each learner (Chondamrongkul et al., 2025). Using vector-based semantic retrieval, it extracts relevant materials from a structured knowledge base. It also implements regression testing and re-indexing to ensure content consistency and relevance.

4.2. Educational content development and assessment (RQ2)

RAG has also been utilized in educational content development and assessment. This area focuses on enhancing the creation and evaluation of educational materials. Different from interactive learning systems, this approach automates content generation without requiring multiple interactions between users and the system.

To address RQ2, this study identifies three major trends in how RAG supports the development and assessment of educational content: (1) Automated Generation of Questions and Lesson Plans, (2) Intelligent Assessment and Feedback Systems, and (3) Knowledge-Based Content Retrieval. These trends demonstrate how RAG enhances the efficiency, accuracy, and contextual relevance of educational material creation and evaluation.

4.2.1. Automated generation of question and lesson plan

Automated content generation in education, particularly in question generation and lesson planning, has been a key area of exploration for RAG. One line of research focuses on leveraging RAG to improve the quality and contextual relevance of generated multiple-choice questions (MCQs). The MCQGen framework integrates LLMs with RAG to retrieve existing MCQs from a curated knowledge base which consists of both teacher-authored and student-created questions (Hang et al., 2024). By using retrieved content as a reference, the model generates new MCQs while ensuring alignment with educational objectives. The study further optimizes prompt engineering techniques, such as Chain-of-Thought (CoT) prompting and self-refinement, to enhance the accuracy and challenge level of the generated questions. A key innovation is the dynamic improvement cycle, where student responses are used to iteratively refine the question set, making it more adaptive to personalized learning needs. Another study examines the feasibility of AI-generated questions in educational settings, introducing ALINet, a system designed to generate questions based on multimodal educational materials, including lecture audio, slides, and reading materials (Zeghouani et al., 2024). MCQGen focuses on structured MCQs, whereas ALINet emphasizes open-ended question generation. It employs an RAG mechanism to retrieve relevant content from lecture materials to mitigate context loss in question generation, thereby improving coherence. Additionally, the study assesses question quality using both automated metrics and subjective evaluations from teachers and students, highlighting the trade-off between automation and human oversight. Beyond question generation, RAG has been applied to the automated generation of lesson plans. One study explores a three-stage pipeline where RAG first retrieves relevant lesson plan components, and then an LLM generates an initial draft, and a self-critique mechanism refines the output (Zheng et al., 2024). This iterative process aims to improve the quality and coherence of generated lesson plans by identifying deficiencies and suggesting improvements. The study applies this approach to elementary mathematics curricula, demonstrating its adaptability across multiple grade levels and topics.

4.2.2. Intelligent assessment and feedback systems

Automated Assessment. The integration of RAG into automated grading has been explored in multiple studies, primarily aiming to enhance grading accuracy, consistency, and explainability. A study introduces SteLLA, a structured scoring system that adopts a RAG approach for structured scoring of short answers (Qiu et al., 2024). It utilizes teacher-provided reference answers and rubrics as an external knowledge base to retrieve relevant information, generate structured scoring criteria, and create evaluation questions. This study assesses student responses through question answering, providing more fine-grained feedback and scoring. Another study evaluated LLMs, including GPT-3.5, GPT-4, Claude-3, and Mistral-Large, in assessing students' open-ended written responses (Jauhiainen & Guerra, 2024). By using RAG to retrieve relevant reference materials, the study aims to mitigate grading variability and improve fairness. The research also systematically examines the impact of model parameters, such as temperature settings, on grading consistency. In the context of evaluating tutoring practices, an approach was proposed to assess teachers' socio-emotional instructional strategies using GPT-3.5 and GPT-4. This study compares various prompting strategies, including zero-shot, Tree of Thought (ToT), and RAG-based approaches, demonstrating that RAG improves grading accuracy by leveraging external knowledge, such as social-emotional learning (SEL) principles (Han et al., 2024).

Learning feedback generation. Beyond grading, RAG has been applied to intelligent feedback generation, particularly in providing personalized and context-aware responses. One study explores a RAG-powered feedback tool integrated into UML modeling software (Ardimento et al., 2024). This system retrieves domain-specific knowledge from a structured vector database and employs LLMs to generate feedback on UML class diagrams, assisting students in refining their modeling practices. Similarly, another study introduces a RAG-enhanced feedback system for K-12 data science education (Fung et al., 2024). By retrieving relevant materials from an educational knowledge base, the system generates explanations tailored to students' cognitive levels. The study compares a RAG-enhanced feedback mechanism with direct GPT-4 generated feedback, finding that RAG improves clarity and alignment with curriculum standards. In the domain of lecture-based learning, a study proposes a RAG-powered feedback system that utilizes lecture transcripts as a retrieval source to enhance the accuracy and relevance of feedback for coding assignments (Jacobs & Jaschke, 2024). By dynamically linking feedback to specific lecture segments with embedded timestamps, students can directly access relevant instructional content, reinforcing learning and reducing hallucinations in AI-generated responses.

4.2.3. Knowledge-based content retrieval

Research in this field has focused on advancing retrieval mechanisms through techniques such as knowledge graph construction, semantic search optimization, and curriculum-aligned IR. One approach emphasizes cross-data knowledge graph construction to enhance LLM-based educational Q&A systems. A study proposes a framework that integrates educational FAQs, course content, and LMS data into a unified knowledge graph, improving the accuracy of intent recognition and policy matching (Bui et al., 2024). The retrieval process is supported by a Neo4j knowledge graph, where Sentence-BERT embeddings and TF-IDF re-ranking refine the retrieval quality before passing the retrieved information to an LLM for final response generation. Another study investigates the retrieval of scientific knowledge to help students understand unfamiliar concepts by developing ScienceDirect Topic Pages, a structured knowledge base of scientific concepts (Capari et al., 2024). The retrieval system utilizes a fine-tuned BERT-based ranking model to extract relevant information from a corpus of over 5.8 million articles, followed by a RAG pipeline to generate structured definitions. Notably, the study incorporates SelfCheckGPT NLI Score to mitigate hallucination issues in retrieved content. In the context of medical education, a study introduces a semantic search engine designed for anatomy textbooks (El Malhi et al., 2024). The system employs CamemBERT-Bio embeddings and the Siamese network with triplet loss to improve retrieval accuracy for French-language medical content. This approach leverages semantic similarity to match student queries with textbook content, rather than relying on traditional keyword-based retrieval. A novel perspective is presented in a study that applies RAG for curriculum-linked retrieval, aiming to automate the alignment of diagnostic tests with course objectives (Li et al., 2024). This system integrates semantic text similarity (STS) retrieval via text-embedding-ada-002 with a classification step using ChatGPT, ensuring that test questions are mapped to the most relevant curriculum goals. The study finds that shorter contextual segments yield better retrieval accuracy, reducing noise in the retrieval phase.

4.3. Scaling RAG in educational ecosystems (RQ3)

The integration of RAG into educational ecosystems is revolutionizing how institutions deliver precise, personalized, and accessible solutions.

This work identifies three major trends in how RAG is scaled and integrated within educational ecosystems for RQ3: (1) AI in Educational Policy and Student Support, (2) Domain-Specific Applications in Various Disciplines, and (3) RAG Enhancement for Language Models and Research Directions. These trends illustrate how RAG has been expanded beyond individual learning systems to broader institutional and disciplinary applications, enabling scalable and context-aware educational support.

4.3.1. AI in educational policy and student support

The application of RAG in education spans multiple stages of the student lifecycle, from university admissions to dropout prediction. In the admissions process, RAG has been employed to handle multi-institutional queries, improving the accuracy and efficiency of university application guidance. A system integrating GPT-3.5 with LlamaIndex, for example, enables more effective responses to complex inquiries related to university application guidance, outperforming traditional FAQ-based systems in both accuracy and user acceptance (Z. Chen et al., 2024). Besides, HICON AI leverages RAG to provide personalized guidance for college applications (Singla et al., 2024). This system categorizes students into distinct advising profiles and integrates resume screening mechanisms to provide personalized recommendations. Notably, unlike conventional text-based chatbots, HICON AI supports voice interactions, improving accessibility for diverse learners. As students progress in their academic journeys beyond admissions, institutions also leverage RAG to address challenges related to student retention. In dropout prediction, RAG has been used alongside few-shot learning to retrieve relevant historical data, offering a more interpretable alternative to traditional machine learning models (Aboukacem et al., 2024). While this approach improves the transparency of predictive insights, challenges related to computational efficiency and data heterogeneity remain.

4.3.2. Domain-specific applications in various disciplines

The application of RAG in education has been explored across various disciplines, offering solutions tailored to domain-specific challenges. In medical education, for instance, retrieval-enhanced AI models have been evaluated for their effectiveness in question-answering tasks. A study comparing a custom GPT model optimized with BM25 retrieval against GPT-4o, which integrates real-time retrieval capabilities, found that retrieval-based approaches provided more reliable responses (Gong et al., 2024). Similarly, in programming and data science education, the use of RAG has gained traction in AI-assisted coding and debugging. Tools like LangChain facilitate more contextually relevant and accurate code generation by retrieving relevant documentation and examples, thereby supporting software development training (Bakharia & Abdi, 2024). Additionally, in geographical and environmental studies, RAG has been proposed as a strategy for mitigating AI-generated hallucinations (Lane, 2025). By allowing educators to curate reliable sources for retrieval, RAG can help address ethical concerns related to bias, data privacy, and accessibility disparities in AI-assisted learning. In music education and vocal training, RAG has been leveraged to address hallucination issues in domain-specific question-answering (Leung et al., 2024). By retrieving authoritative sources, RAG enhances the accuracy of AI-generated responses, making it a valuable tool for disciplines where contextual precision is crucial.

4.3.3. RAG enhancement for language models and research directions

The application of RAG in education extends beyond individual tutoring systems and automated content generation. Several studies have explored its broader implications, including its role in large-scale online education, educational policy analysis, and the integration of AI into subject-specific learning. One significant area of research is the use of RAG in large-scale digital education and intelligent learning environments. In MOOCs, RAG-enhanced GPT models have been shown to improve response accuracy in assessments, particularly across different question types (Miladi et al., 2024b). In a related study, researchers explored whether small language models (SLMs) equipped with RAG could be viable alternatives to LLM in computer science education (Liu et al., 2024c). This study also introduced an Indexed Context (IC) framework, which employs a multi-layered retrieval structure to ensure that the retrieved educational content aligns with students' learning needs. Their findings indicated that an optimized SLM combined with RAG could deliver performance comparable to LLMs. Additionally, research on the engineering challenges of RAG systems has identified key failure points in AI-driven educational assistants, emphasizing the need for robust retrieval mechanisms and continuous monitoring to ensure reliable performance (Barnett et al., 2024).

5. Challenges and future direction

RAG has demonstrated significant potential in enhancing educational applications by improving the accuracy and contextual relevance of AI-generated responses. However, several challenges persist, limiting its effectiveness in real-world educational settings. This section examines these challenges and explores potential directions for future improvements.

5.1. Hallucinated content

Although the retrieval mechanism helps improve factual accuracy by incorporating external knowledge, it cannot entirely prevent the model from generating inaccurate or misleading responses (Fernandez et al., 2024; Ehrenthal et al., 2024; Capari et al., 2024). This limitation can be attributed to several underlying factors. First, the retrieval process itself may surface documents that are loosely related to the query or even contain outdated or misleading information (Yu et al., 2024; Baek et al., 2023). Second, even when relevant documents are retrieved, they may lack sufficient depth or completeness of information, leaving gaps that the language model attempts to fill, sometimes inaccurately (Fernandez et al., 2024). Third, inaccuracies may also exist in human-annotated ground-truth answers. For instance, SyllabusQA examines the reliability of human-annotated answers and identifies several potential sources of error (Fernandez et al., 2024). The primary factors contributing to such errors include imperfect recall, annotator mistakes, ambiguities within the syllabus itself, and inherent ambiguities in the answers.

Recent real-world deployments highlight the severity of hallucinated content in educational contexts. For example, Harvard's CS50 course integrated a RAG-enhanced AI assistant to support students in understanding programming concepts (R. Liu et al., 2024). Although the assistant retrieved information from curated lecture materials, it occasionally exhibited hallucinations, confidently delivering incorrect explanations. Acknowledging the risks posed to student learning, the course mandated that all AI-generated answers be reviewed and endorsed by human instructors, emphasizing the continued necessity of human oversight in RAG-augmented educational environments.

Although RAG has implemented mitigation strategies, hallucinated content remains a critical issue. Students often lack the expertise to critically evaluate AI-generated answers, making them vulnerable to internalizing misinformation, which can hinder their learning and cognitive development. This limitation also prevents educational chatbots and Q&A systems from fully replacing human teachers. Although approaches like self-checking mechanisms (Dan et al., 2023) and iterative hallucination detection (Capari et al., 2024) have been proposed, they are not yet sufficiently effective to ensure reliable outputs without human supervision.

Possible solutions. Future research should focus on improving hallucination detection by integrating methods for uncertainty estimation or human-in-the-loop validation to enhance response reliability. In practical terms, educators are encouraged to implement a double-verification workflow in which AI-generated outputs undergo systematic human review before being utilized in instructional materials or assessments. Faculty training initiatives focused on developing critical evaluation skills for AI responses could further enhance content reliability in classrooms. From the policy perspective, there is a need to establish mandatory guidelines that require human oversight of AI-generated educational content. Policymakers could also consider developing certification frameworks that validate the factual accuracy and transparency standards of RAG-based educational tools before they are deployed at scale.

5.2. *Completeness and timeliness of RAG knowledge base*

Since RAG retrieves information from external repositories, the accuracy and relevance of its responses are directly influenced by the quality of these sources. However, most existing RAG systems rely on static databases that do not update automatically to incorporate newly available knowledge. This limitation is particularly problematic in educational settings, where knowledge spans multiple domains and evolves continuously.

Recent studies, such as SKYRAG, have highlighted the critical importance of timely knowledge updates in educational RAG applications (Soekamto et al., 2025). SKYRAG addresses this challenge by dynamically retrieving course information from multiple MOOC databases, ensuring that learning paths remain aligned with newly added or updated courses. Without such dynamic retrieval, learners risk being guided by outdated curricula, particularly in fast-evolving fields like computer science and data science. This underscores the necessity for RAG systems to incorporate mechanisms for real-time knowledge base updates in education.

Possible solutions. To address this issue, future developments should focus on integrating real-time retrieval from authoritative sources, enhancing automated data ingestion pipelines, and incorporating mechanisms to assess the reliability and relevance of retrieved documents before generating responses. One promising direction is the adoption of agentic RAG, which introduces autonomous agents capable of actively monitoring, retrieving, and validating new information (Singh et al., 2025). Besides, educators should take an active role in maintaining and updating institutional knowledge repositories, ensuring that the underlying databases used by RAG systems are aligned with the most recent developments in each discipline. Curating course-specific supplemental resources can further enhance retrieval quality. Policymakers, on the other hand, should support the development of centralized, continuously updated educational repositories accessible to multiple institutions, thereby promoting equitable access to high-quality, up-to-date knowledge bases.

5.3. *High cost and latency in RAG*

Currently, the generation stage of RAG primarily relies on LLM-based APIs, with costs driven by two key factors: the high price of API tokens and the lengthy inference time of LLMs. First, many educational RAG systems depend on commercial LLM APIs, which result in large costs, particularly in educational applications where frequent queries are required for student support and assessment (Yan et al., 2024; Neumann et al., 2025; Han et al., 2024). For instance, Neumann et al. (2025) estimated that the cost per student could reach approximately $1.65 USD, while Han et al. (2024) compared API pricing across different models, highlighting the financial burden on education-focused RAG applications. Second, the inherent complexity of retrieval and generation leads to high inference latency, which restricts the scalability of RAG in real-world deployments (Jauhiainen & Guerra, 2024; Aboukacem et al., 2024; Arun et al., 2024). Specifically, Jauhiainen and Guerra (2024) analyzed the average inference speeds of various models, demonstrating the variations in response times. Moreover, Arun et al. (2024) showed that although GPT models can generate responses almost instantaneously, certain proposed methods require up to 15 seconds, which may hinder real-time interactions in educational settings.

Possible solutions. To address these challenges, future research should explore cost-efficient alternatives, such as lightweight open-source models and optimized retrieval mechanisms, to mitigate these limitations while maintaining accuracy. Educators are advised to optimize the structure and complexity of AI queries to reduce token usage and operational costs and consider the adoption of cost-effective, institution-hosted models where feasible. By selecting more efficient RAG configurations, educational institutions can make AI technologies more sustainable and scalable. Policymakers should consider providing financial support to institutions for deploying private or locally hosted RAG infrastructures and should encourage the development and adoption of affordable AI solutions specifically tailored to educational needs.

5.4. *Limited multimodal support in RAG-based educational tools*

Many disciplines, such as medical education, engineering, and the visual arts, rely heavily on non-textual resources, including images, program code (Dakshit, 2024), videos (Z. Chen et al., 2024), and interactive simulations (Izquierdo-Domenech et al., 2024). However, most current RAG models primarily process and generate text, making them less effective in these fields (Hennekeuser et al., 2024; Ehrenthal et al., 2024). The faculty feedback highlights this issue, noting that RAG struggles with images and equations (Dakshit, 2024), which are essential for interpreting engineering schematics, analyzing pathology slides, or understanding artistic compositions.

In real-world educational settings, these limitations have led to significant challenges. For example, faculty members observed that RAG systems failed to accurately process graphical content and mathematical formulas embedded in computer science course materials (Dakshit, 2024). Similarly, in university admissions support, the inability to handle visual elements, such as application flowcharts and infographics, restricted the chatbot's ability to deliver comprehensive information to prospective students (Z. Chen et al., 2024). These shortcomings critically constrain the broader applicability of RAG systems in domains where multimodal understanding is indispensable.

Possible solutions. Future advancements should focus on integrating vision-language models or cross-modal retrieval techniques to enable more comprehensive content generation. Additionally, improving the alignment between multimodal resources and textual explanations could enhance the learning experience by providing richer, more contextually relevant educational materials. In the meantime, educators should proactively supplement AI-generated outputs with curated visual, auditory, and interactive resources to better support disciplines that rely on non-textual materials, such as engineering, medicine, and the arts. Policymakers can facilitate progress in this area by funding research initiatives dedicated to multimodal RAG development and by promoting standards that ensure educational AI systems are capable of processing diverse content formats crucial for comprehensive learning.

6. Conclusion

In this paper, we provide a comprehensive review of RAG in the educational domain, focusing on both its technical foundations and real-world applications. We first introduce the principles of RAG and explain its three key stages: Indexing, Retrieval, and Generation. This framework enhances knowledge retrieval and improves response reliability in AI-driven educational systems. We then examine its diverse applications across various educational contexts, including interactive learning systems, educational content development and assessment, and broader educational ecosystems. These implementations demonstrate RAG's potential to support personalized learning, reduce teachers' workload, and address the limitations of traditional LLM in a variety of domains.

Despite its advantages, RAG faces several challenges, including hallucination, difficulties in integrating updated knowledge, high computational costs, and limited multimodal support. Overcoming these issues requires improving retrieval efficiency, incorporating additional validation mechanisms, implementing dynamic knowledge updates, and expanding multimodal retrieval capabilities. By integrating generative AI with external knowledge retrieval, RAG offers a promising direction for AI-driven education, enabling more accurate, context-aware, and adaptive learning experiences.

CRediT authorship contribution statement

Zongxi Li: Writing – original draft, Supervision, Conceptualization. **Zijian Wang:** Writing – original draft, Investigation, Formal analysis. **Weiming Wang:** Writing – review & editing, Funding acquisition, Conceptualization. **Kevin Hung:** Writing – review & editing, Resources, Funding acquisition. **Haoran Xie:** Writing – review & editing, Supervision, Project administration. **Fu Lee Wang:** Writing – review & editing, Supervision.

Declaration of competing interest

The author, Prof Xie Haoran, is an Editor-in-Chief for this journal and was not involved in the editorial review or the decision to publish this article.

Acknowledgement

This work has been supported by a grant from Hong Kong Metropolitan University (Project Reference No. CP/2022/02), by a Faculty Research Grant (SDS24A2) from Lingnan University, by the Hong Kong Research Grants Council through the Faculty Development Scheme (Project No. UGC/FDS16/E10/23), and by the Teaching Development Grant entitled "Flipping and Gamifying Data Science Classrooms" (102722) of Lingnan University, Hong Kong.

References

Aboukacem, A., Berrada, I., Bergou, E. H., Iraqi, Y., & Mekouar, L. (2024). Investigating the predictive potential of large language models in student dropout prediction. In *International conference on artificial intelligence in education* (pp. 381–388). Springer.

Alier, M., Pereira, J., García-Peñalvo, F. J., Casañ, M. J., & Cabré, J. (2025). Lamb: An open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems. *Computer Standards & Interfaces*, *92*, Article 103940. https://doi.org/10.1016/j.csi.2024.103940. https://www.sciencedirect.com/science/article/pii/S0920548924001090.

Alon, U., Xu, F., He, J., Sengupta, S., Roth, D., & Neubig, G. (2022). Neuro-symbolic language modeling with automaton-augmented retrieval. In *International conference on machine learning* (pp. 468–485).

Ardimento, P., Bernardi, M. L., Cimitile, M., & Scalera, M. (2024). A rag-based feedback tool to augment uml class diagram learning. In *Proceedings of the ACM/IEEE 27th international conference on model driven engineering languages and systems* (pp. 26–30). Association for Computing Machinery.

Arun, G., Perumal, V., Urias, F. P. J. B., Ler, Y. E., Tan, B. W. T., Vallabhajosyula, R., Tan, E., Ng, O., Ng, K. B., & Mogali, S. R. (2024). Chatgpt versus a customized AI chatbot (anatbuddy) for anatomy education: A comparative pilot study. *Anatomical Sciences Education*, *17*, 1396–1405.

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The twelfth international conference on learning representations*.

Baek, J., Jeong, S., Kang, M., Park, J. C., & Hwang, S. J. (2023). Knowledge-augmented language model verification. In *The 2023 conference on empirical methods in natural language processing*. https://openreview.net/forum?id=sOngusZCsN.

Bakharia, A., & Abdi, S. (2024). Shaping programming and data science education: Insights from genai technical book trends. In *2024 IEEE international conference on advanced learning technologies (ICALT)* (pp. 116–120). IEEE.

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd international conference on AI engineering-software engineering for AI* (pp. 194–199).

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J. B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. In *International conference on machine learning* (pp. 2206–2240).

Bui, T., Tran, O., Nguyen, P., Ho, B., Nguyen, L., Bui, T., & Quan, T. (2024). Cross-data knowledge graph construction for llm-enabled educational question-answering system: A case study at hcmut. In *Proceedings of the 1st ACM workshop on AI-powered Q&A systems for multimedia* (pp. 36–43).

Capari, A., Azarbonyad, H., Tsatsaronis, G., Afzal, Z., & Dunham, J. (2024). Sciencedirect topic pages: A knowledge base of scientific concepts across various science domains. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval* (pp. 2976–2980).

Chen, K. F., Hwang, G. J., & Chen, M. R. A. (2024). Effects of a concept mapping-guided virtual laboratory learning approach on students' science process skills and behavioral patterns. *Educational Technology Research and Development*, *72*, 1623–1651.

Chen, X., Xie, H., & Hwang, G. J. (2020). A multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. *Computers and Education: Artificial Intelligence*, *1*, Article 100005.

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *1*, Article 100002.

Chen, X., Xie, H., Li, Z., Cheng, G., Leng, M., & Wang, F. L. (2023). Information fusion and artificial intelligence for smart healthcare: A bibliometric study. *Information Processing & Management*, *60*, Article 103113.

Chen, Z., Zou, D., Xie, H., Lou, H., & Pang, Z. (2024). Facilitating university admission using a chatbot based on large language models with retrieval-augmented generation. *Educational Technology & Society*, *27*, 454–470.

Cheng, X., Luo, D., Chen, X., Liu, L., Zhao, D., & Yan, R. (2023). Lift yourself up: Retrieval-augmented text generation with self-memory. In *Thirty-seventh conference on neural information processing systems*. https://openreview.net/forum?id=lYNSvp51a7.

Chondamrongkul, N., Hristov, G., & Temdee, P. (2025). Addressing technical challenges in large language model-driven educational software system. *IEEE Access*, *13*, 12846–12858. https://doi.org/10.1109/ACCESS.2025.3531380.

Dakshit, S. (2024). Faculty perspectives on the potential of rag in computer science higher education. In *Proceedings of the 25th annual conference on information technology education* (pp. 19–24).

Dan, Y., Lei, Z., Gu, Y., Li, Y., Yin, J., Lin, J., Ye, L., Tie, Z., Zhou, Y., Wang, Y., et al. (2023). Educhat: A large-scale language model-based chatbot system for intelligent education. arXiv preprint, arXiv:2308.02773.

Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, *57*, 1–39.

Dehbozorgi, N., Kunuku, M. T., & Pouriyeh, S. (2024). Personalized pedagogy through a llm-based recommender system. In *International conference on artificial intelligence in education* (pp. 63–70). Springer.

Dong, C. (2023). How to build an AI tutor that can adapt to any course and provide accurate answers using large language model and retrieval-augmented generation. arXiv preprint, arXiv:2311.17696.

Ehrenthal, J. C., Gachnang, P., Loran, L., Rahms, H., & Schenker, F. (2024). Integrating generative artificial intelligence into supply chain management education using the scor model. In *International conference on advanced information systems engineering* (pp. 59–71). Springer.

El Malhi, M., Talbi, M., Lamti, S., & Kerzazi, N. (2024). Semantic search engine within anatomy books: A bert-based model for medical students. In *International conference on smart medical, IoT & artificial intelligence* (pp. 108–115). Springer.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T. S., & Li, Q. (2024). A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 6491–6501).

Fernandez, N., Scarlatos, A., & Lan, A. (2024). SyllabusQA: A course logistics question answering dataset. In *Long papers: Vol. 1. Proceedings of the 62nd annual meeting of the Association for Computational Linguistics* (pp. 10344–10369). Association for Computational Linguistics.

Fung, S. C. E., Wong, M. F., & Tan, C. W. (2024). Automatic feedback generation on k-12 students' data science education by prompting cloud-based large language models. In *Proceedings of the eleventh ACM conference on learning@ scale* (pp. 255–258).

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint, arXiv:2312.10997.

Glass, M., Rossiello, G., Chowdhury, M. F. M., Naik, A., Cai, P., & Gliozzo, A. (2022). Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies, association for computational linguistics: Human language technologies* (pp. 2701–2715).

Gong, E. J., Bang, C. S., Lee, J. J., Park, J., Kim, E., Kim, S., Kimm, M., & Choi, S. H. (2024). The potential clinical utility of the customized large language model in gastroenterology: A pilot study. *Bioengineering*, *12*, 1.

Guo, K., & Li, D. (2024). Understanding efl students' use of self-made AI chatbots as personalized writing assistance tools: A mixed methods study. *System*, *124*, Article 103362. https://doi.org/10.1016/j.system.2024.103362. https://www.sciencedirect.com/science/article/pii/S0346251X24001441.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. In *International conference on machine learning* (pp. 3929–3938).

Han, Z. F., Lin, J., Gurung, A., Thomas, D., Chen, E., Borchers, C., Gupta, S., & Koedinger, K. (2024). Improving assessment of tutoring practices using retrieval-augmented generation. In *AI for education: Bridging innovation and responsibility at the 38th AAAI annual conference on AI*. https://openreview.net/forum?id=Us1EF795Ys.

Hang, C. N., Wei, Tan C., & Yu, P. D. (2024). Mcqgen: A large language model-driven mcq generator for personalized learning. *IEEE Access*, *12*, 102261–102273. https://doi.org/10.1109/ACCESS.2024.3420709.

Henkel, O., Levonian, Z., Li, C., & Postle, M. (2024). Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. In *Proceedings of the 17th international conference on educational data mining* (pp. 315–320). International Educational Data Mining Society.

Hennekeuser, D., Vaziri, D. D., Golchinfar, D., Schreiber, D., & Stevens, G. (2024). Enlarged education–exploring the use of generative AI to support lecturing in higher education. *International Journal of Artificial Intelligence in Education*, 1–33.

Hu, W., Tian, J., & Li, Y. (2025). Enhancing student engagement in online collaborative writing through a generative AI-based conversational agent. *The Internet and Higher Education*, *65*, Article 100979. https://doi.org/10.1016/j.iheduc.2024.100979. https://www.sciencedirect.com/science/article/pii/S1096751624000411.

Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education.

Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume* (pp. 874–880).

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2022). Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=jKN1pXi7b0.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, *24*, 1–43.

Izquierdo-Domenech, J., Linares-Pellicer, J., & Ferri-Molla, I. (2024). Virtual reality and language models, a new frontier in learning. *International Journal Of Interactive Multimedia and Artificial Intelligence*, *8*. https://doi.org/10.9781/ijimai.2024.02.007.

Jacobs, S., & Jaschke, S. (2024). Leveraging lecture content for improved feedback: Explorations with gpt-4 and retrieval augmented generation. In *2024 36th international conference on software engineering education and training (CSEE&T)* (pp. 1–5). IEEE.

Jauhiainen, J. S., & Guerra, A. G. (2024). Evaluating students' open-ended written responses with llms: Using the rag framework for gpt-3.5, gpt-4, claude-3, and mistral-large. *Advances in Artificial Intelligence and Machine Learning*, *4*, 3097–3113.

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, *28*, 15873–15892.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*. https://doi.org/10.1145/3571730.

Ji, Z., Liu, Z., Lee, N., Yu, T., Wilie, B., Zeng, M., & Fung, P. (2023). RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the association for computational linguistics: ACL 2023* (pp. 4504–4522).

Jiang, Z., Xu, F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 7969–7992).

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, *7*, 535–547.

Karpukhin, V., Oguz, B., Min, S., Lewis, P. S., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. In *EMNLP: Vol. 1* (pp. 6769–6781).

Ko, H. T., Liu, Y. K., Tsai, Y. C., & Suen, S. (2024). Enhancing python learning through retrieval-augmented generation: A theoretical and applied innovation in generative AI education. In *Innovative technologies and learning* (pp. 164–173). Switzerland: Springer Nature.

Lane, R. (2025). Mitigating risks, embracing potential: A framework for integrating generative artificial intelligence in geographical and environmental education. *International Research in Geographical and Environmental Education*. https://doi.org/10.1080/10382046.2025.2458561.

Lee, Y. (2024). Developing a computer-based tutor utilizing generative artificial intelligence (gai) and retrieval-augmented generation (rag). *Education and Information Technologies*, 1–22.

Leung, C. h. J., Yi, Y., Kuai, L., Li, Z., Yeung, S. k. A., Lee, K. w. J., Ho, K. h. K., & Hung, K. (2024). Rag for question-answering for vocal training based on domain knowledge base. In *2024 11th international conference on behavioral and social computing (BESC)* (pp. 1–6).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Li, X., Henriksson, A., Duneld, M., Nouri, J., & Wu, Y. (2024). Supporting teaching-to-the-curriculum by linking diagnostic tests to curriculum goals: Using textbook content as context for retrieval-augmented generation with large language models. In *International conference on artificial intelligence in education* (pp. 118–132). Springer.

Li, X., Li, Z., Li, J., Xie, H., & Li, Q. (2025). ESE: Espresso sentence embeddings. In *The thirteenth international conference on learning representations*. https://openreview.net/forum?id=plgLA2YBLH.

Li, Z., Li, Y., Xie, H., & Qin, S. J. (2025). Condambigqa: A benchmark and dataset for conditional ambiguous question answering. arXiv preprint, arXiv:2502.01523.

Liang, H. Y., Hwang, G. J., Hsu, T. Y., & Yeh, J. Y. (2024). Effect of an AI-based chatbot on students' learning performance in alternate reality game-based museum learning. *British Journal of Educational Technology*, *55*, 2315–2338.

Liu, C., Hoang, L., Stolman, A., & Wu, B. (2024). Hita: A rag-based educational platform that centers educators in the instructional loop. In *Artificial intelligence in education* (pp. 405–412). Switzerland: Springer Nature.

Liu, R., Zenke, C., Liu, C., Holmes, A., Thornton, P., & Malan, D. J. (2024). *Teaching cs50 with AI: Leveraging generative artificial intelligence in computer science educationIn Proceedings of the 55th ACM technical symposium on computer science education: Vol. 1* (pp. 750–756). Association for Computing Machinery.

Liu, S., Yu, Z., Huang, F., Bulbulia, Y., Bergen, A., & Liut, M. (2024c). Can small language models with retrieval-augmented generation replace large language models when learning computer science? In *Proceedings of the 2024 on innovation and technology in computer science education: Vol. 1* (pp. 388–393).

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Long papers: Vol. 1. Proceedings of the 61st annual meeting of the Association for Computational Linguistics* (pp. 9802–9822).

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, *35*, 17359–17372.

Miladi, F., Psyché, V., & Lemire, D. (2024a). Comparative performance of gpt-4, rag-augmented gpt-4, and students in moocs. In A. Basiouni, & C. Frasson (Eds.), *Breaking barriers with generative intelligence. Using GI to improve human education and well-being* (pp. 81–92). Switzerland: Springer Nature.

Miladi, F., Psyché, V., & Lemire, D. (2024b). Leveraging gpt-4 for accuracy in education: A comparative study on retrieval-augmented generation in moocs. In *International conference on artificial intelligence in education* (pp. 427–434). Springer.

Miladi, F., Psyché, V., Diattara, A., El Mawas, N., & Lemire, D. (2025). Evaluating a gpt-4 and retrieval-augmented generation-based conversational agent to enhance learning experience in a mooc. In *Proceedings of the international conference on computer supported education*.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M. A., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. arXiv:2402.06196.

Neumann, A. T., Yin, Y., Sowe, S., Decker, S., & Jarke, M. (2025). An llm-driven chatbot in higher education for databases and information systems. *IEEE Transactions on Education*, *68*, 103–116. https://doi.org/10.1109/TE.2024.3467912.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, Article 100041.

Prihar, E., Lee, M., Hopman, M., Kalai, A. T., Vempala, S., Wang, A., Wickline, G., Murray, A., & Heffernan, N. (2023). Comparing different approaches to generating mathematics explanations using large language models. In *International conference on artificial intelligence in education* (pp. 290–295). Springer.

Qiu, H., White, B., Ding, A., Costa, R., Hachem, A., Ding, W., & Chen, P. (2024). Stella: A structured grading system using llms with rag. In *2024 IEEE international conference on big data (BigData)* (pp. 8154–8163). IEEE.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023a). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, *11*, 1316–1331.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023b). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, *11*, 1316–1331. https://doi.org/10.1162/tacl_a_00605. https://aclanthology.org/2023.tacl-1.75/.

Rao, J., & Lin, J. (2024). Ramo: Retrieval-augmented generation for enhancing moocs recommendations. arXiv preprint, arXiv:2407.04925.

Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, *3*, 333–389.

Roemer, R. C., & Borchardt, R. (2015). Altmetrics and the role of librarians. *Library Technology Reports*, *51*, 31–37.

Sachan, D. S., Patwary, M., Shoeybi, M., Kant, N., Ping, W., Hamilton, W. L., & Catanzaro, B. (2021). End-to-end training of neural retrievers for open-domain question answering. arXiv preprint, arXiv:2101.00408.

Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., & Zaharia, M. (2022). ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 3715–3734).

Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. T. (2025). Agentic retrieval-augmented generation: A survey on agentic rag. arXiv preprint, arXiv:2501.09136.

Singla, A. D., Tripathi, S., & Victoria, A. H. (2024). Hicon AI: Higher education counseling bot. In *2024 4th international conference on pervasive computing and social networking (ICPCSN)* (pp. 779–784). IEEE.

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, *11*, 1–17. https://doi.org/10.1162/tacl_a_00530.

Soekamto, Y. S., Limanjaya, L. C., Purwanto, Y. K., & Kang, D. K. (2025). From queries to courses: Skyrag's revolution in learning path generation via keyword-based document retrieval. *IEEE Access*, *13*, 21434–21455. https://doi.org/10.1109/ACCESS.2025.3535618.

Soliman, H., Kravcik, M., Neumann, A. T., Yin, Y., Pengel, N., & Haag, M. (2024). Scalable mentoring support with a large language model chatbot. In *European conference on technology enhanced learning* (pp. 260–266). Springer.

Stockwell, G. (2022). *Mobile assisted language learning concepts, contexts and challenges*. Cambridge University Press. Copyright: © Glenn Stockwell 2022.

Taneja, K., Maiti, P., Kakar, S., Guruprasad, P., Rao, S., & Goel, A. K. (2024). Jill Watson: A virtual teaching assistant powered by chatgpt. In *Artificial intelligence in education* (pp. 324–337). Switzerland: Springer Nature.

Teng, D., Wang, X., Xia, Y., Zhang, Y., Tang, L., Chen, Q., Zhang, R., Xie, S., & Yu, W. (2024). Investigating the utilization and impact of large language model-based intelligent teaching assistants in flipped classrooms. *Education and Information Technologies*, 1–34.

Thüs, D., Malone, S., & Brünken, R. (2024). Exploring generative AI in higher education: A rag system to enhance student engagement with scientific literature. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1474892.

Wang, K., Ramos, J., & Lawrence, R. (2024). Chated: A chatbot leveraging chatgpt for an enhanced learning experience in higher education. In *INTED2024 proceedings* (pp. 6580–6589). IATED.

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., & Wen, Q. (2024). Large language models for education: A survey and outlook. arXiv preprint, arXiv:2403.18105.

Xu, F., Shi, W., & Choi, E. (2024). RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The twelfth international conference on learning representations*. https://openreview.net/forum?id=mlJLVigNHp.

Yan, L., Zhao, L., Echeverria, V., Jin, Y., Alfredo, R., Li, X., Gaševi'c, D., & Martinez-Maldonado, R. (2024). Vizchat: Enhancing learning analytics dashboards with contextualized explanations using multimodal generative AI chatbots. In *International conference on artificial intelligence in education* (pp. 180–193). Springer.

Yu, W., Zhang, H., Pan, X., Cao, P., Ma, K., Li, J., Wang, H., & Yu, D. (2024). Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 14672–14685).

Zeghouani, O., Ali, Z., Simson van Dijkhuizen, W., Hong, J. W., & Clos, J. (2024). Examining the feasibility of AI-generated questions in educational settings. In *Proceedings of the second international symposium on trustworthy autonomous systems*. Association for Computing Machinery.

Zhang, N., Yao, Y., Tian, B., Wang, P., Deng, S., Wang, M., Xi, Z., Mao, S., Zhang, J., Ni, Y., et al. (2024). A comprehensive study of knowledge editing for large language models. arXiv preprint, arXiv:2401.01286.

Zhang, Y., He, R., Liu, Z., Lim, K. H., & Bing, L. (2020). An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 1601–1610).

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*, 1–38.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. arXiv preprint, arXiv:2303.18223.

Zheng, Y., Li, X., Huang, Y., Liang, Q., Guo, T., Hou, M., Gao, B., Tian, M., Liu, Z., & Luo, W. (2024). Automatic lesson plan generation via large language models with self-critique prompting. In *International conference on artificial intelligence in education* (pp. 163–178). Springer.