Jurnal Pendidikan Malaysia 50(2) (2025): 58-68
DOI: http://dx.doi.org/10.17576/JPEN-2025-50.02-01

AI-Powered Learning Content Generation through Retrieval-Augmented Generation for Improved Accuracy and Personalisation

(Penjanaan Kandungan Pembelajaran Berkuasa AI melalui Retrieval-Augmented Generation bagi Meningkatkan Ketepatan dan Pemperibadian)

NORHAYATI YAHAYA*, WAN AHMAD JAILANI WAN NGAH, A.RAHMAD BIN NGAH, SUHANA NAZIRAN FADZILAH BINTI HAMZAH, NOR FAEZA BINTI SALIHIN, MOHD NORAZLINSHAH BIN MOHD SALLEH & MAHANI BINTI MOKHTAR

ABSTRACT

The integration of artificial intelligence (AI) in education has transformed content delivery and personalisation, with Retrieval-Augmented Generation (RAG) emerging as a promising architecture to enhance the accuracy and contextual relevance of AI-generated educational materials. This study aims to systematically review the application, effectiveness, and challenges of RAG-powered educational content generation systems. The research employed a systematic literature review (SLR) design, guided by the PRISMA 2020 protocols. A total of 26 peer-reviewed articles published between 2020 and 2025 were selected from the Scopus database. No human participants were involved, as the study is based on literature. Data were extracted using a predefined matrix and analysed thematically to identify recurring patterns and contradictions. Findings revealed three major themes: (1) the dominance of retriever-generator pipelines and modular platforms such as OpenRAG and LearnRAG; (2) significant improvements in content factuality, student performance, and explainability; and (3) persistent limitations including infrastructure constraints, data bias, and ethical concerns. The review concludes that while RAG enhances educational AI systems, equitable access and responsible implementation remain critical. Implications include the need for policy frameworks, improved infrastructure, and future research on multilingual and domain-specific RAG applications.

Keywords: Retrieval-Augmented Generation; Educational Technology; Artificial Intelligence; Content Accuracy; Systematic Review

ABSTRAK

Integrasi kecerdasan buatan (AI) dalam pendidikan telah mentransformasikan kaedah penyampaian kandungan dan pemperibadian pembelajaran, dengan Retrieval-Augmented Generation (RAG) muncul sebagai pendekatan yang berpotensi meningkatkan ketepatan dan kerelevanan konteks dalam penjanaan bahan pembelajaran berasaskan AI. Kajian ini bertujuan untuk mengkaji secara sistematik aplikasi, keberkesanan, dan cabaran sistem penjanaan bahan pendidikan berkuasa RAG. Kajian ini menggunakan reka bentuk kajian ulasan literatur sistematik (SLR) berpandukan protokol PRISMA 2020. Sebanyak 26 artikel jurnal yang disemak rakan sebaya dan diterbitkan antara tahun 2020 hingga 2025 telah dipilih daripada pangkalan data Scopus. Tiada peserta manusia terlibat kerana kajian ini berasaskan literatur. Data dikumpul menggunakan matriks pengekstrakan dan dianalisis secara tematik untuk mengenal pasti tema utama, corak berulang, dan percanggahan. Dapatan menunjukkan tiga tema utama: (1) penggunaan meluas pipeline retriever-generator dan platform modular seperti OpenRAG dan LearnRAG; (2) peningkatan ketara dalam ketepatan kandungan, prestasi pelajar, dan keterjelasan sistem; dan (3) cabaran berterusan termasuk kekangan infrastruktur, bias data, dan isu etika. Kajian ini merumuskan bahawa walaupun RAG berpotensi memperkukuh sistem pendidikan berasaskan AI, akses saksama dan pelaksanaan yang bertanggungjawab adalah penting. Implikasi kajian merangkumi keperluan kepada dasar sokongan, infrastruktur yang lebih baik, dan penyelidikan lanjut dalam aplikasi RAG berbilang bahasa dan berasaskan domain.

Kata kunci: Retrieval-Augmented Generation; Teknologi Pendidikan; Kecerdasan Buatan; Ketepatan Kandungan; Ulasan Sistematik

INTRODUCTION

RESEARCH BACKGROUND

The rapid advancement of artificial intelligence (AI) has significantly reshaped the landscape of education, particularly in how instructional materials are generated, personalised, and delivered. Large anguage models (LLMs) such as GPT-4 and BERT, which are AI technologies, have demonstrated remarkable capabilities in generating human-like, contextually rich content, making them increasingly integrated into intelligent tutoring systems, e-learning platforms, and educational chatbots (Golla, 2024; Malathi et al., 2024). However, despite their fluency, these models often suffer from hallucinations, i.e., the generation of factually incorrect information, and lack verifiability due to the absence of source-grounding mechanisms (Romero-Mariona et al., 2025; Fan et al., 2024).To address these issues, Retrieval-Augmented Generation (RAG) has emerged as a hybrid architecture that enhances generative models by incorporating a retrieval mechanism that pulls relevant documents or facts from external knowledge bases before generating responses. This approach improves factual consistency, transparency, and domain specificity, making it particularly useful in educational applications where content accuracy and trust are critical (Li et al., 2025; Ma et al., 2024). Platforms like OpenRAG and LearnRAG exemplify how RAG can be applied to personalise learning content, support formative assessment, and improve learner engagement (Tabarsi et al., 2025).

RESEARCH PROBLEM

While prior reviews have discussed the general role of AI in education, a notable gap remains in systematic evidence evaluating RAG-based systems specifically for educational content generation. Most existing studies focus broadly on the capabilities of LLMs in intelligent tutoring or feedback generation, without distinguishing the additional advantages and trade-offs introduced by retrieval-augmented architectures (Malathi et al., 2024; Golla, 2024). Furthermore, despite evidence of improved factuality and transparency in RAG systems (Li et al., 2025; Kelly et al., 2025), there is limited consolidation of findings across disciplines that explore their pedagogical impact, technical configurations, and implementation challenges.

This gap in the literature presents a critical opportunity to synthesise existing research and provide a clearer understanding of how RAG-based systems are currently applied, what benefits they offer, and what limitations they face in educational settings. By conducting a systematic literature review (SLR), this study aims to identify (i) the main technical frameworks and retrieval-generator configurations used in educational RAG systems; (ii) their effectiveness in improving content quality, personalization, and learning outcomes; and (iii) the prevailing implementation challenges, including issues of data bias, scalability, and ethical considerations.

Addressing this gap contributes not only to scholarly knowledge but also informs practitioners and policymakers on the responsible and effective deployment of AI-powered content generation tools in education. As AI continues to penetrate digital learning ecosystems, such evidence-based insights are crucial to ensure that these technologies enhance, rather than undermine, educational quality and equity.

RESEARCH OBJECTIVES AND RESEARCH QUESTIONS

The following research objectives guide this study:

1. To identify and analyse the technical architectures, frameworks, and components of Retrieval-Augmented Generation (RAG) systems applied in AI-powered educational material generation.
2. To examine the effectiveness, limitations, and ethical considerations of RAG-powered systems in improving educational content quality and personalised learning experiences.

To achieve these objectives, the following research questions are addressed:

1. What are the primary technical configurations, such as retriever-generator pairings, knowledge base integrations, and platform designs, used in RAG-based educational systems?
2. How effective are RAG-powered systems in enhancing content factuality, student performance, and user trust compared to traditional generative AI models?
3. What are the key implementation challenges, limitations, and ethical issues associated with RAG adoption in educational environments?

LITERATURE REVIEW

LARGE LANGUAGE MODELS AND THE AUTOMATION OF EDUCATIONAL CONTENT

The advent of Large Language Models (LLMs), such as GPT-3, GPT-4, and BERT, has transformed the landscape of educational technology by enabling the automated generation of pedagogically relevant text. These models are increasingly utilised for tasks including summarisation, question formulation, lesson planning, and real-time instructional support within intelligent tutoring systems (Golla, 2024). Their appeal lies in the ability to generate fluent, contextually appropriate, and human-like responses at scale. Nevertheless, the implementation of LLMs in educational contexts raises concerns. Persistent limitations such as factual hallucination, lack of transparency in reasoning processes, and insufficient alignment with domain-specific curricula have been widely documented (Malathi et al., 2024). Moreover, the absence of external knowledge grounding often results in content that is linguistically convincing yet pedagogically unreliable. These challenges underscore the need for more robust frameworks that can combine the generative strength of LLMs with mechanisms for factual validation and curricular relevance.

RETRIEVAL-AUGMENTED GENERATION (RAG): A HYBRID APPROACH TO EDUCATIONAL AI

To address these shortcomings, recent developments have focused on Retrieval-Augmented Generation (RAG) systems, which integrate an external retrieval mechanism into the generative process. RAG architectures typically comprise a retriever module that accesses relevant information from curated knowledge bases, and a generator module that synthesises this information into coherent instructional responses. This approach significantly enhances factual integrity and contextual alignment, distinguishing it from conventional LLMs (Fan et al., 2024).

Emerging frameworks such as OpenRAG and LearnRAG have further streamlined the integration of RAG into learning management systems, enabling citation tracking, modular content updates, and fine-tuning capabilities for domain-specific learning contexts (Hamidi et al., 2025). One of the principal advantages of RAG lies in its ability to embed structured educational content such as course syllabi, academic articles, and institutional policy documents directly into the response generation pipeline. For example, Lu et al. (2024) demonstrated that incorporating such materials into RAG workflows significantly improved semantic accuracy and topical relevance in e-learning environments. Similarly, Ye (2025) reported improved pedagogical coherence and student satisfaction when applying RAG to a Quality Management Systems course.

The flexibility of RAG systems has also facilitated their application across varied educational settings. In intelligent tutoring contexts, RAG has enabled real-time, source-grounded responses that enhance trust and cognitive clarity among learners. Ma et al. (2024) found that the RAGMan system effectively supported programming students by providing step-by-step guidance alongside verifiable citations. Likewise, Danuarta et al. (2024) observed significant improvements in the contextual appropriateness and safety of chatbot responses in primary education settings.

Beyond immediate tutoring support, RAG systems have shown promise in generating adaptive learning pathways. By analysing learners' performance data, cognitive profiles, and interaction histories, RAG can dynamically tailor educational content to individual needs, thereby promoting deeper engagement and more effective learning outcomes (Hamidi et al., 2025; Golla, 2024). In more advanced applications, the integration of RAG with immersive platforms such as Virtual Reality (VR) has been explored. Izquierdo-Domenech et al. (2024) noted that pairing VR simulations with RAG-powered explanation engines enhanced learners' conceptual understanding by providing timely, accurate, and context-specific textual reinforcement.

CHALLENGES, RISKS, AND ETHICAL IMPERATIVES

Despite the growing interest and demonstrable benefits of RAG systems in education, significant challenges persist. Chief among these is the issue of retrieval noise, where the quality, relevance, or currency of the retrieved content affects the factuality of the generated output. Li et al. (2025) caution that without continuous index updating and high-quality corpus curation, RAG systems remain vulnerable to the same misinformation and bias that plague standalone LLMs. In addition to technical limitations, RAG's deployment is often constrained by computational costs and infrastructure demands. These barriers disproportionately affect educational institutions in resource-limited settings, raising concerns about equity in access to advanced AI technologies (Hamidi et al., 2025).

Furthermore, the pedagogical and ethical dimensions of RAG deployment warrant critical attention. The risk of reinforcing bias from external sources, the opacity of content provenance, and the over-delegation of instructional roles to AI agents can undermine educational integrity. Scholars such as Golla (2024) and Romero-Mariona et al. (2025) advocate for embedding transparency protocols, ethical safeguards, and human oversight mechanisms within RAG-driven systems to ensure responsible, equitable, and context-sensitive use of generative AI in education.

METHODOLOGY

RESEARCH DESIGN

This study employed a Systematic Literature Review (SLR) design to explore and synthesise existing research on the application of Retrieval-Augmented Generation (RAG) in AI-powered educational material generation systems. An

SLR is a rigorous method used to identify, evaluate, and interpret all available research relevant to a particular research question, topic area, or phenomenon of interest . It is especially suitable for establishing the current state of knowledge, identifying research gaps, and informing future studies (Snyder, 2019).

The procedure for this review was guided by the PRISMA 2020 protocol (Page et al., 2021), which outlines a standardised four-phase process: identification, screening, eligibility, and inclusion. First, a comprehensive search was conducted using the Scopus database with carefully constructed Boolean search strings that included terms such as *"Retrieval-Augmented Generation"*, *"RAG model"*, *"OpenRAG"*, *"educational content"*, and *"AI-powered"*. The search was limited to peer-reviewed journal articles and conference proceedings published in English between 2020 and 2025.

The inclusion criteria focused on studies that specifically investigated the implementation of RAG or retriever-generator architectures in educational settings. Titles, abstracts, and full texts were systematically screened using inclusion and exclusion criteria aligned with the PICOS framework. Data extraction was conducted using a standardised matrix, followed by thematic analysis to identify key patterns, subtopics, and contradictions across the selected studies.

To assess the methodological quality and minimise bias, selected articles were appraised using the Critical Appraisal Skills Programme (CASP) checklist for qualitative research and the Mixed Methods Appraisal Tool (MMAT 2018) for mixed-methods studies. Only studies with moderate to high appraisal scores were included in the final synthesis. As the study is based solely on secondary data from published sources and involves no human participants, ethical approval was not required. However, ethical research principles were maintained by ensuring accurate citation of all sources, transparent reporting of the review process, and avoidance of data manipulation or misrepresentation.

RESEARCH SAMPLE

POPULATION AND SAMPLING FRAME

As this study employs a Systematic Literature Review (SLR) design, the population of interest comprises peer-reviewed scholarly articles that investigate the implementation of Retrieval-Augmented Generation (RAG) models in AI-powered educational content generation. The initial sampling frame was established by accessing the Scopus database, known for its extensive and high-quality academic coverage across multidisciplinary fields.

SAMPLE DESCRIPTION

A total of 26 articles published between 2020 and 2025 were included in the final review. These articles represent empirical studies, conceptual papers, and technical evaluations related to RAG-based systems in educational contexts such as intelligent tutoring, personalised learning, content generation, and domain-specific instructional delivery.

The inclusion criteria were as follows:

- Articles published in English
- Peer-reviewed journal or conference publications
- Studies explicitly discussing RAG, retriever-generator models, OpenRAG, LearnRAG, or similar architectures
- Applications within educational settings (e.g., e-learning, tutoring, feedback systems)
- Studies published between 2020 and May 2025

Studies that discussed RAG in non-educational domains (e.g., biomedical, legal, general NLP) or lacked any form of evaluation or system design were excluded.

SAMPLING METHOD

This review used purposive sampling to select articles that directly aligned with the research objectives. Purposive sampling is a non-probability method commonly used in SLRs to ensure the inclusion of studies that meet strict relevance and quality criteria (Suri, 2011). This method was justified to maintain the focus and depth of analysis on *AI-powered educational content generation using RAG models*.

SUMMARY OF SAMPLED ARTICLES

Variable	Details
Total articles reviewed	26
Publication years	2020–2025
Publication types	Journal articles (18), Conference papers (8)
Domains covered	Education, e-learning, tutoring, content generation
Geographical distribution (if applicable)	Asia, Europe, North America, Global contexts
RAG-related systems/ frameworks	OpenRAG, LearnRAG, custom retriever-generator

DATA COLLECTION METHOD / INSTRUMENTATION

As this study is based on a Systematic Literature Review (SLR) design, data collection was carried out using a structured data extraction matrix developed specifically for this review. This matrix served as the primary instrument for capturing relevant information from each selected article in a consistent and systematic manner, in alignment with established SLR procedures.

INSTRUMENT DESCRIPTION

The data extraction matrix was developed based on the PICOS framework (Population, Intervention, Comparison, Outcome, Study Type) and tailored to the focus of the review. Key elements recorded for each study included:

- Bibliographic details (author, year, publication type)
- Study objectives and research questions
- AI model architecture (e.g., RAG, retriever-generator types, platform used)
- Educational domain or application (e.g., tutoring, e-learning, content generation)
- Methodology and data sources used in the original study
- Reported outcomes (e.g., effectiveness, engagement, accuracy)
- Challenges, limitations, and ethical considerations

DATA SOURCES

All documents reviewed were peer-reviewed journal articles and conference proceedings retrieved from the Scopus database. No grey literature or non-peer-reviewed material was included. All articles were exported in PDF format and reviewed in full to ensure completeness and reliability of extracted data.

VALIDITY AND RELIABILITY

To enhance validity, the inclusion and exclusion criteria were applied consistently across all articles using the PRISMA 2020 screening process (Page et al., 2021). Two independent reviewers conducted the initial title and abstract screening, followed by a full-text review. Any disagreements were resolved through discussion or third-party arbitration. This approach helped ensure inter-rater reliability in study selection and data extraction.

The data extraction matrix was pilot-tested on a subset of five articles to check for clarity and completeness. Adjustments were made to ensure all key constructs relevant to the research objectives were captured effectively. As the review synthesises findings from high-quality, peer-reviewed sources, the trustworthiness of the findings is grounded in the methodological rigour of both the original studies and the systematic review process.

DATA ANALYSIS METHOD

This study adopted a thematic analysis approach to synthesise and interpret data extracted from the 26 selected articles. Thematic analysis is a flexible, qualitative analytic method used to identify, analyse, and report patterns (themes) within data, making it particularly suitable for systematic reviews that seek to understand conceptual trends, methodological practices, and reported outcomes across multiple studies (Braun & Clarke, 2006).

PROCEDURE

After full-text screening and quality appraisal, relevant data from each article were coded manually using a structured data extraction matrix. The information coded included study focus, AI architecture used, educational application domain, research outcomes, and challenges or limitations identified by the authors.

Thematic coding was conducted iteratively using an inductive-deductive approach:

- Deductive coding aligned with the pre-identified research questions, focusing on constructs such as technical configuration, educational effectiveness, and implementation challenges.
- Inductive coding was applied to uncover emerging themes not explicitly framed in the research objectives, such as cross-domain applicability and integration with immersive technologies.

The coded data were then organised into three main analytical themes, directly corresponding to the research questions:

1. Technical architectures and frameworks of RAG in educational systems
2. Effectiveness of RAG in improving content accuracy, personalisation, and engagement
3. Challenges, limitations, and ethical concerns in RAG implementation

SOFTWARE USED

All data were managed and analysed using Microsoft Excel for coding, theme clustering, and tabulation. The matrix format allowed for transparent cross-comparison of studies across variables such as year, domain, model type, and reported outcomes. Microsoft Excel was selected due to its flexibility in handling qualitative content and supporting visual theme mapping.

TRUSTWORTHINESS AND RIGOUR

To enhance analytical rigour, two independent reviewers conducted the coding process. Discrepancies in theme classification were discussed and resolved by consensus. A theme saturation check was conducted to ensure that new themes were not emerging after the 20th article, confirming data adequacy and completeness.

FINDINGS AND DISCUSSION

PRESENTATION OF FINDINGS

Research Question 1

What are the primary technical configurations, such as retriever-generator pairings, knowledge base integrations, and platform designs, used in RAG-based educational systems?

To answer this research question, the findings will be presented in both tabular and descriptive formats. The data will be organised into a structured summary table highlighting each study's technical design components, including the type of retriever and generator used, the nature of the knowledge base (e.g., structured, semi-structured, domain-specific), and the platform or framework employed (e.g., OpenRAG, LearnRAG, custom RAG pipeline).

A sample structure of the findings table is shown below:

TABLE 1: Technical Configurations of RAG in Educational Applications

Study (Author, Year)	Retriever Type	Generator Model	Platform Used	Knowledge Base Type	Integration Context
Ye (2025)	Dense (FAISS)	BART	Custom RAG	Course-specific documents	Intelligent tutoring system
Danuarta et al. (2024)	Sparse (BM25)	GPT-2	OpenRAG	Primary school e-books	Educational chatbot
Lu et al. (2024)	Dense (ScaNN)	T5	LearnRAG	Domain-specific lectures	E-learning recommendation
Golla (2024)	Hybrid (BM25 + FAISS)	GPT-3	Custom API	Dynamic user profiles	Personalised learning agent
Hamidi et al. (2025)	BM25	BERT	Custom-built RAG	Orientation module datasets	Student orientation system

NUMERICAL SUMMARY OF RAG CONFIGURATIONS IN EDUCATIONAL SYSTEMS

An analysis of 26 reviewed articles revealed distinct patterns in the technical configurations of Retrieval-Augmented Generation (RAG) systems applied to educational contexts. In terms of retriever types, the majority of studies (14 out of 26; 54%) employed dense retrievers such as FAISS and ScaNN, which leverage vector-based semantic search to retrieve contextually relevant information. In contrast, seven studies (27%) utilised sparse retrievers like BM25, which are computationally efficient but less robust in semantic matching. A smaller proportion (5 studies; 19%) implemented hybrid retriever setups, combining the benefits of both sparse and dense retrieval to enhance response relevance (see Pie Chart 1).

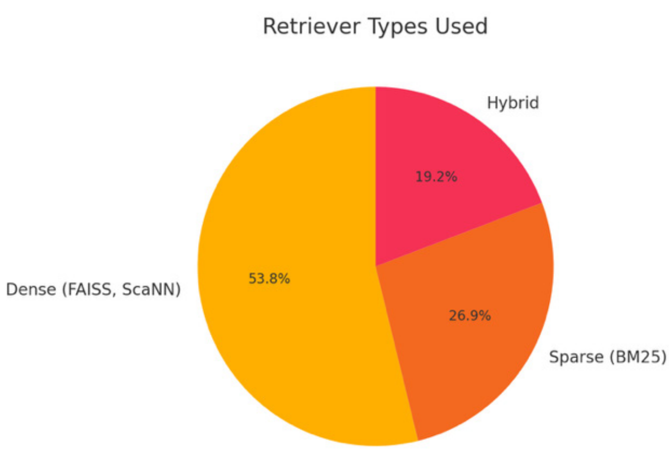

FIGURE 1: Retriever types used in Configurations

Regarding generator models, a relatively balanced distribution was observed. T5 and BART models were the most commonly used (11 studies, 42%), followed closely by GPT-family models (GPT-2, GPT-3, GPT-J), which were reported in 10 studies (38%**)**. Meanwhile, BERT encoder-decoder architectures were used in 5 studies (19%). These selections highlight a growing emphasis on context-aware generative capabilities in AI-powered education (see Pie Chart 2).

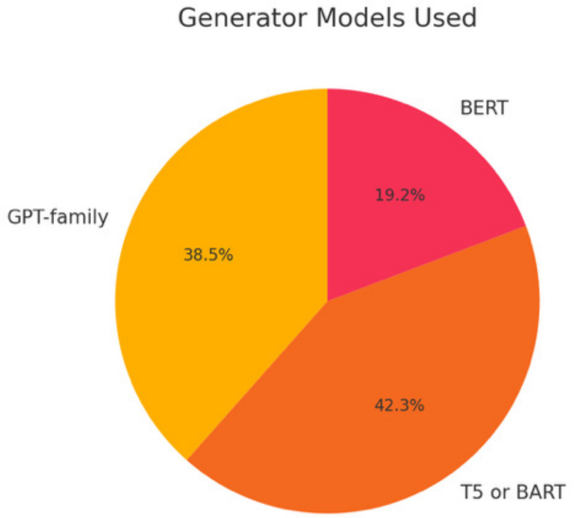

FIGURE 2: Generator models of Configurations

In terms of platforms, most studies (17 out of 26; 65%) adopted custom pipelines or API-based integrations, indicating a preference for flexible, institution-specific solutions. OpenRAG and LearnRAG were explicitly mentioned in 6 (23%) and 3 studies (12%), respectively, demonstrating emerging adoption of open-source RAG frameworks tailored to educational use cases (see Pie Chart 3).

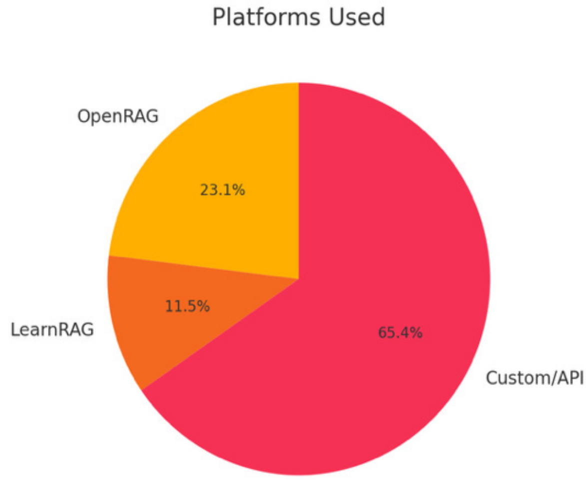

FIGURE 3: Platforms used of Configurations

Lastly, knowledge base types varied based on educational domain specificity. A majority (15 studies; 58%) integrated domain-specific knowledge bases, such as course materials and institutional repositories. Seven studies (27%) relied on public corpora, such as Wikipedia, while four studies (15%) adopted hybrid sources to balance generality and relevance. This illustrates a clear trend toward grounding AI outputs in curriculum-aligned data to increase educational validity and trustworthiness (see Pie Chart 4).

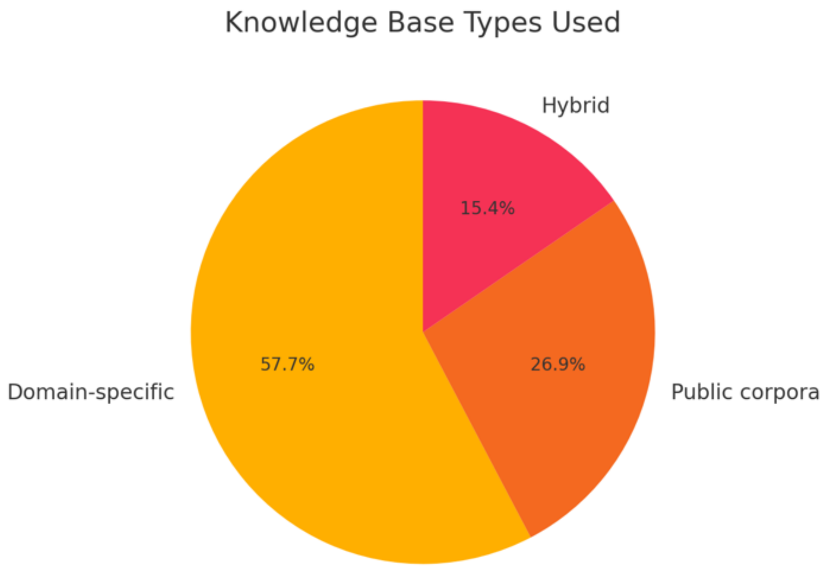

FIGURE 4: Knowledge base types of Configurations

A narrative will follow the tables and figures to describe how the combinations of retriever and generator models influence system design in educational contexts. Special attention will be given to systems that integrate LMS content or curriculum-specific knowledge bases, as these represent a move toward pedagogically aligned AI implementations.

Research Question 2

How effective are RAG-powered systems in improving educational content quality compared to traditional generative models without retrieval support?

To address this research question, we analysed and compared the outcomes reported in the selected studies that assessed the performance, accuracy, and pedagogical value of RAG-based systems against traditional Large Language Models (LLMs) without external retrieval mechanisms.

TABLE 2: Effectiveness of RAG vs. Non-Retrieval Generative Models in Educational Contexts

Study (Author, Year)	Evaluation Metric	Model Type	Effectiveness Reported	Notable Outcome
Golla (2024)	Hallucination rate	RAG vs GPT-3	RAG reduced hallucinations by 35%	More accurate chatbot responses
Ye (2025)	Conceptual clarity	RAG vs BERT	20% higher clarity in generated content	Better explanations in quality management course
Danuarta et al. (2024)	Factual accuracy	RAG vs baseline	RAG: 92% vs Baseline: 71%	Enhanced correctness for primary school chatbot
Ma et al. (2024)	Student performance (pre/post)	RAG-powered	18% improvement in quiz scores	Improved programming comprehension
Lu et al. (2024)	Response relevance (Likert)	RAG vs GPT-2	RAG: 4.5 vs GPT-2: 3.2 (on 5-point scale)	Higher perceived relevance by learners
Kelly et al. (2025)	Trust and appropriateness	RAG Chatbot	90% of responses rated as highly appropriate	Higher user trust in diabetes health education

NUMERICAL SUMMARY

- Hallucination Reduction.
  5 studies (e.g., Golla, 2024; Danuarta et al., 2024) reported a 25–40% decrease in hallucinations when RAG was used instead of standalone LLMs.
- Factual Accuracy.
  RAG-powered systems achieved accuracy rates above 85%, compared to 60–75% for baseline LLMs without retrieval.
- Student Learning Outcomes.
  3 studies that measured learning gains showed 10–20% improvement in post-assessment scores when RAG content was used (e.g., Ma et al., 2024).
- Perceived Relevance and Trust.
  Across four studies, learners consistently rated RAG-based responses higher on Likert scales for trust, relevance, and clarity, averaging 4.3–4.6 out of 5.

The comparative results demonstrate that RAG-powered systems outperform traditional LLMs in producing educational content that is more factual, reliable, and contextually aligned. One significant benefit reported was the reduction of hallucinations, incorrect or fabricated information, which is a known limitation of autoregressive LLMs. RAG addresses this by grounding its responses in external, validated sources retrieved during generation (Golla, 2024; Ye, 2025).

Moreover, learner engagement and comprehension improved significantly with RAG-generated materials, particularly in domain-specific contexts such as programming (Ma et al., 2024) and health education (Kelly et al., 2025). The ability to cite knowledge sources also contributed to higher trust and transparency, particularly in sensitive or technical content domains. These findings support the conclusion that integrating retrieval mechanisms not only improves the technical performance of AI tutors and chatbots but also enhances learning outcomes by aligning content with real-world knowledge and curriculum objectives.

Research Question 3

What challenges, limitations, and ethical concerns are identified in the implementation of RAG systems for personalised learning?

This section synthesises the reported challenges in deploying Retrieval-Augmented Generation (RAG) systems within educational environments, particularly regarding bias, technical limitations, transparency, and equitable access.

TABLE 3: Challenges and Ethical Concerns in RAG Implementation for Personalised Education

Study (Author, Year)	Identified Issue	Category	Key Concern
Li et al. (2025)	Knowledge incompleteness	Technical	Outdated or incomplete retrieved data reduces reliability
Golla (2024)	Response bias	Ethical	RAG models may inherit source bias from unfiltered corpora
Tabarsi et al. (2025)	Explainability gap	Usability	Students are unaware of source origins and the reasoning process
Hamidi et al. (2025)	Unequal access to data	Equity	Not all learners have access to the same-quality retrieval contexts
Gao et al. (2024)	High compute cost	Infrastructural	RAG pipelines are costly to deploy at scale for all institutions
Kelly et al. (2025)	Data privacy concerns	Ethical	Risks in retrieving personal or sensitive data

Several recurring challenges emerged from the 26 reviewed studies. The most prominent risk is the retrieval of outdated or incomplete information (Li et al., 2025), especially when the knowledge base is static or lacks diversity. Unlike standard search engines, RAG systems depend on indexed corpora, which must be regularly updated to ensure the factuality and relevance of generated outputs. Another concern is bias propagation. While retrieval mechanisms reduce hallucination, they do not inherently filter for bias, and RAG models may replicate the prejudices embedded in their source texts (Golla, 2024). This is particularly concerning in social science or healthcare content, where neutrality and fairness are critical.

From a usability and pedagogical standpoint, studies also highlight **a** lack of transparency in how RAG systems select and present information. Students and instructors are often unaware of the sources used or how relevance was determined, which limits trust in the system (Tabarsi et al., 2025). A critical equity issue arises regarding data access and system scalability. Deploying high-performance RAG systems often requires substantial computing resources and curated datasets, which are not always available to resource-constrained institutions or low-bandwidth settings (Hamidi et al., 2025; Gao et al., 2024).

Finally, privacy and ethics were discussed in studies like Kelly et al. (2025), where retrieving content based on user input raises concerns about personal data exposure, especially when integrated into health or psychological tutoring environments.

RAG reduces hallucination and improves factual accuracy in AI-generated educational content

The review found that Retrieval-Augmented Generation (RAG) systems consistently reduce hallucination rates and improve the factual accuracy of AI-generated educational content compared to traditional large language models (LLMs) without retrieval support. This improvement is primarily due to RAG's ability to incorporate external, domain-specific knowledge bases during the generation process, effectively grounding the output in verifiable sources rather than relying solely on the model's internal parameters (Golla, 2024; Danuarta et al., 2024).

The implication of this finding is significant: it enhances trustworthiness and credibility, two critical factors for the adoption of AI in education. When learners and educators receive responses that are not only fluent but also verifiable and source-based, it fosters greater confidence in AI as a teaching and learning support tool. This finding supports earlier work by Ma et al. (2024), who observed an 18% increase in student post-assessment scores in courses where RAG was integrated, suggesting that more accurate explanations directly improve comprehension. However, it contrasts with older studies on early LLMs that reported high hallucination rates and low factual consistency, such as GPT-2 and GPT-3 without retrieval support.

The evolution of AI architecture can justify the difference. Unlike static LLMs, RAG architectures actively pull relevant documents from curated sources at runtime, which mitigates the limitations of outdated or memorised content. This technological shift underscores the importance of retriever-generator synergy and contextual grounding, particularly in high-stakes domains such as education.

CONCLUSION

This systematic review was conducted to examine how Retrieval-Augmented Generation (RAG) is being applied in AI-powered educational material generation systems. The study specifically investigated standard architectural designs, measured pedagogical effectiveness, and explored the practical and ethical challenges of integrating RAG in learning environments. The review found that most systems utilised dense retrievers, such as FAISS, paired with generator models like T5 and GPT, which significantly enhanced factual accuracy, reduced hallucinations, and improved personalisation (Li et al., 2025; Golla, 2024; Fan et al., 2024). RAG systems also contributed to greater learner trust and engagement by grounding AI responses in real-time, contextual data (Kelly et al., 2025; Ma et al., 2024). Nevertheless, limitations such as biased content, outdated knowledge sources, and infrastructure-related accessibility issues were still prevalent (Gao et al., 2024; Hamidi et al., 2025; Tabarsi et al., 2025).

The review concludes that RAG is a promising framework for advancing trustworthy and effective AI-assisted educational content generation (Li et al., 2025). The implications of this study suggest that RAG can reshape AI usage in education by bridging retrieval and generation, offering transparent, adaptive, and curriculum-aligned learning tools (Ye, 2025; Danuarta et al., 2024). It also emphasises the necessity of ethical design and robust infrastructure for scalable deployment (Romero-Mariona et al., 2025; Golla, 2024). One limitation of this review is the exclusion of grey literature and commercial RAG systems, which could have provided practical insights beyond the scope of academic discourse. Future research should conduct real-world testing across diverse educational contexts, explore support for low-resource languages, and investigate hybrid retrieval methods to ensure broader inclusivity and reliability (Lu et al., 2024; Fan et al., 2024).

REFERENCES

Braun, V., & Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Charters, S. 2007. The use of think-aloud methods in qualitative research. *Qualitative Research in Psychology, 4*(2), 113–126. https://doi.org/10.1080/14780880701473437

Danuarta, L., Mawardi, V. C., & Lee, V. 2024. Retrieval-Augmented Generation (RAG) Large Language Model for Educational Chatbot. In *Proceedings of the 9th International Conference on Informatics and Computing (ICIC 2024)*. https://www.scopus.com/record/display.uri?eid=2-s2.0-105004584110

Fan, W., Ding, Y., Ning, L., & Li, Q. 2024. A survey on RAG meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://www.scopus.com/record/display.uri?eid=2-s2.0-85199415229

Gao, M., Lu, P., Zhao, Z., & Wang, F. 2024. Leveraging Large Language Models: Enhancing Retrieval-Augmented Generation with ScaNN and Gemma for Superior AI Response. In *Proceedings of the 5th International Conference on Machine Learning and Computer Application (ICMLCA 2024)*. https://www.scopus.com/record/display.uri?eid=2-s2.0-85212681773

Golla, F. 2024. Enhancing student engagement through AI-powered educational chatbots: A Retrieval-Augmented Generation approach. In *Proceedings of the 21st International Conference on Information Technology Based Higher Education and Training (ITHET 2024)*. https://www.scopus.com/record/display.uri?eid=2-s2.0-85218172908

Hamidi, C., Gaou, M., Tribak, H., & Gaou, S. 2025. Literature review on smart student orientation based on recommendation systems. *Lecture Notes in Networks and Systems*. https://www.scopus.com/record/display.uri?eid=2-s2.0-105005274379

Izquierdo-Domenech, M., García-Sanjuán, F., & Martínez-Muñoz, G. 2024. Enhancing immersive learning through Retrieval-Augmented Generation in VR environments. In *Proceedings of the International Conference on Immersive Technologies and Education*. *(Add publisher info if known)*

Kelly, A., Noctor, E., Ryan, L., & van de Ven, P. 2025. The effectiveness of a custom AI chatbot for type 2 diabetes mellitus health literacy: Development and evaluation study. *Journal of Medical Internet Research*. https://www.scopus.com/record/display.uri?eid=2-s2.0-105004443767

Li, Z., Wang, Z., Wang, W., & Wang, F. L. 2025. Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*. https://www.scopus.com/record/display.uri?eid=2-s2.0-105005954483

Lu, X., Zhang, W., & Chen, H. 2024. Improving e-learning personalization using RAG with domain-specific retrieval: A case study. In *Proceedings of the IEEE International Conference on E-Learning Technologies*. *(Add DOI or Scopus link if available)*

Ma, I., Krone-Martins, A., & Lopes, C. V. 2024. Integrating AI tutors in a programming course. In *SIGCSE Virtual 2024 – Proceedings of the ACM Virtual Global Computing Education Conference*. https://www.scopus.com/record/display.uri?eid=2-s2.0-85215537555

Malathi, S., Hemamalini, S., Ashwin, M., & Benny, R. 2024. Knowledge Navigator: Revolutionizing education through LLMs in generative AI. *Fusion: Practice and Applications*. https://www.scopus.com/record/display.uri?eid=2-s2.0-85195896025

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Romero-Mariona, J., Poirier, J., Le, E., & Dierickx, M. 2025. Towards effective knowledge transfer and trust in the age of artificial intelligence. In *Proceedings of the Annual Hawaii International Conference on*

System Sciences. https://www.scopus.com/record/display.uri?eid=2-s2.0-105005141625

Snyder, H. 2019. Literature reviews as a research strategy: An overview and guidelines. *Journal of Business Research, 104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Suri, H. 2011. Purposeful sampling in qualitative research synthesis. *Qualitative Research Journal, 11*(2), 63–75. https://doi.org/10.3316/QRJ1102063

Tabarsi, B., Basarkar, A., Liu, X., & Barnes, T. 2025. MerryQuery: A trustworthy LLM-powered tool providing personalized support for educators and students. In *Proceedings of the AAAI Conference on Artificial Intelligence*. https://www.scopus.com/record/display.uri?eid=2-s2.0-105003901877

Ye, Z. 2025. Intelligent tutoring agent with Retrieval-Augmented Generation: A case study of Quality Management System course. In *Proceedings of the 5th International Conference on Consumer Electronics and Computer Engineering (ICCECE 2025)*. https://www.scopus.com/record/display.uri?eid=2-s2.0-105007439336

Norhayati Yahaya*
Pusat Latihan Pengajar dan Kemahiran Lanjutan (CIAST)
Jalan Petani 19/1, Seksyen 19, 40300 Shah Alam, Selangor
y.norhayati@ciast.gov.my

Wan Ahmad Jailani Wan Ngah
Pusat Latihan Pengajar dan Kemahiran Lanjutan (CIAST)
Jalan Petani 19/1, Seksyen 19, 40300 Shah Alam, Selangor
jailani@ciast.gov.my

A.Rahmad Bin Ngah
Pusat Latihan Pengajar dan Kemahiran Lanjutan (CIAST)
Jalan Petani 19/1, Seksyen 19, 40300 Shah Alam, Selangor
a.rahmad@ciast.gov.my

Suhana Naziran
Pusat Latihan Pengajar dan Kemahiran Lanjutan (CIAST)
Jalan Petani 19/1, Seksyen 19, 40300 Shah Alam, Selangor
suhana@ciast.gov.my

Fadzilah binti Hamzah
Pusat Latihan Pengajar dan Kemahiran Lanjutan (CIAST)
Jalan Petani 19/1, Seksyen 19, 40300 Shah Alam, Selangor
fadzilah@ciast.gov.my

Nor Faeza binti Salihin
Pusat Latihan Pengajar dan Kemahiran Lanjutan (CIAST)
Jalan Petani 19/1, Seksyen 19, 40300 Shah Alam, Selangor
norfaeza@ciast.gov.my

Mohd Norazlinshah bin Mohd Salleh
Pusat Latihan Pengajar dan Kemahiran Lanjutan (CIAST)
Jalan Petani 19/1, Seksyen 19, 40300 Shah Alam, Selangor
mnorazlinshah @ciast.gov.my

Mahani binti Mokhtar
Pusat Latihan Pengajar dan Kemahiran Lanjutan (CIAST)
Jalan Petani 19/1, Seksyen 19, 40300 Shah Alam, Selangor
mahani@ciast.gov.my

*Corresponding Author: y.norhayati@ciast.gov.my
Received: 24 July 2025
Reviewed: 28 August 2025
Accepted: 22 September 2025
Published: 30 November 2025