GenEdu: Personalized Learning with Retrieval-Augmented Generative AI and Dual-Layer Evaluation

Roua Frigui[1] and Kalthoum Rezgui[1,2] [1] [2]

Abstract

Personalized learning through generative artificial intelligence (GenAI) is challenged by a fundamental trade-off between adaptability and reliability. Although large language models (LLMs) have opened new possibilities for generating personalized and context-aware educational content, their tendency to produce inaccurate or unsupported information (referred to as hallucinations) remains a major challenge in educational settings. To address this issue, we propose `GenEdu`, a retrieval-augmented generation (RAG)-based AI framework that combines four key learning support features: adaptive quiz generation, level-aware explanations, document-grounded analysis, and educational diagram generation. By integrating RAG with a dual-layer evaluation mechanism, GenEdu delivers personalized learning experiences while ensuring that generated content remains accurate, reliable, and aligned with trusted educational sources. Results from an ablation study reveal that the incorporation of RAG improves semantic alignment by 23.9% according to BERTScore, increases faithfulness by 344.5%, and reduces hallucination rates by 60.8%. In addition to these automatic evaluation metrics, a user-centered study involving 106 participants reports high overall satisfaction ($M = 4.28/5,\ \alpha = 0.952$) across all evaluated dimensions, indicating that grounded generation can simultaneously achieve adaptability and trustworthiness. These results suggest that GenAI can be a reliable tool for education when it is properly designed and carefully evaluated, rather than a source of misleading or incorrect information.

Keywords—Personalized learning, Retrieval-augmented generation, Generative AI, reliability, LLM evaluation.

1 Introduction

Personalized learning seeks to adapt educational content, feedback, pacing, and support strategies to the needs of individual learners. Within Technology-Enhanced Learning (TEL), these objectives have motivated decades of work on adaptive systems, learner modeling, and intelligent tutoring. However, despite substantial progress, many digital learning tools still operate with a one-size-fits-all logic in which the same content and explanations are delivered to all users, regardless of their background, current level, or learning trajectory [1, 2].

The recent rise of LLMs has renewed interest in AI-supported personalization. LLM-based systems can generate explanations, quizzes, hints, and feedback dynamically, thereby enabling more flexible and interactive educational experiences [3, 4]. At the same time, these models introduce critical challenges. In educational contexts, fluent but unsupported AI-generated outputs can mislead learners, reduce trust, and undermine pedagogical validity. Consequently, effective educational use of GenAI requires architectures that combine adaptability with grounding, and generation quality with evaluation rigor [5, 6, 7].

To address this challenge, this paper introduces `GenEdu`, a GenAI framework for personalized learning support. GenEdu was designed as a unified system that integrates four complementary functionalities: (i) Adaptive quiz generation, (ii) Personalized explanation generation, (iii) PDF grounded assistance using RAG, and (iv) AI-assisted diagram generation. Instead of implementing these capabilities as independent modules, the framework unifies them within a coherent learning workflow in which learner interactions and retrieval-based content grounding collaboratively guide the generation process. The main contributions of this work are fourfold. First, an integrated framework combining LLM-based generation with retrieval-augmented mechanisms for personalized educational assistance is introduced. Second, a unified Web-based platform incorporating four pedagogically significant functionalities is developed. Third, a hybrid evaluation methodology that combines user-centered assessment with automatic semantic evaluation metrics is proposed. Finally, empirical evidence is provided showing that retrieval grounding substantially enhances the reliability, faithfulness, and educational value of generated educational content. The remainder of this paper is organized as follows. Section 2 reviews related work and identifies research gaps. Section 3 describes GenEdu's system architecture and evaluation methodology. Section 4 presents results in three layers: ablation study (isolating RAG impact), automatic semantic evaluation (assessing reliability), and user-centered evaluation (assessing usability). Section 5 concludes with limitations and future work.

2 Related work

The integration of GenAI into education has rapidly evolved over the past two years, moving from exploratory chatbot applications to more structured approaches for personalized content delivery, adaptive assessment, and learning support [3, 5, 8].

2.1 LLM-enhanced personalization without grounding

Recent studies have explored the integration of LLMs into personalized learning environments. These approaches leverage learner profiles, engagement data, and performance indicators to guide LLMs in generating adaptive educational content, including proficiency-level explanations, personalized topic recommendations, and customized feedback on learner assignments. Several studies report improvements in flexibility and perceived relevance when generative components are integrated into learning workflows [1, 2, 9]. Yet this body of work reveals a critical gap: most systems rely on the parametric knowledge of the LLM alone, without retrieval grounding. In this approach, the system prompts the model with learner context and relies on the model's internal training knowledge to generate content. While this is flexible, it carries two serious risks. First, the model may hallucinate, generating plausible, sounding but factually incorrect information, contradicting course materials or spreading misconceptions. Second, the model has no mechanism to ensure outputs align with instructor-provided resources, course-specific definitions, or the specific scope of material covered in a course. In education, this misalignment is not merely a quality issue; it is a pedagogical failure.

2.2 Retrieval-augmented educational systems

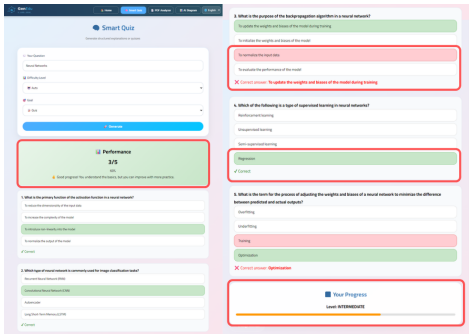
A second major research direction focuses on RAG as a solution to grounding. RAG systems retrieve relevant documents or passages from a knowledge base before generating output, thereby constraining the generation to align with external and verifiable sources. In educational contexts, this grounding is particularly important: generated content must remain aligned with course materials, domain knowledge, and instructor intent. Recent educational applications have explored retrieval-enhanced tutoring, contextual explanation generation, and grounded conversational support [10, 4, 11]. A growing body of evidence suggests that RAG reduces hallucination [6, 7] and improves factual accuracy in generative systems. However, existing educational RAG systems typically address a single task (e.g., tutoring or document analysis or quiz generation) rather than integrating multiple pedagogical functions. Moreover, evaluation has often remained limited in scope, focusing mainly on user feedback or superficial NLP metrics, such as BLEU and ROUGE, while largely overlooking deeper aspects like hallucination, factual consistency, and reliability measures specific to RAG systems.

2.3 Hybrid pedagogical assistants

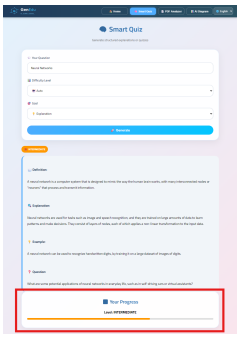
A third research direction involves hybrid systems that combine multiple capabilities, such as tutoring, feedback generation, adaptive assessment, and explanation generation. This design is grounded in pedagogy and reflects the idea that students learn better in a coherent and unified environment where different forms of support are available together, including explanations when they are stuck, quizzes to evaluate understanding, visual aids for conceptual reasoning, and document-based reference material, all integrated within a single continuous interaction. Some prior work has explored such multi-function assistants, but several limitations persist. First, many systems remain specialized, addressing only one or two functions rather than a comprehensive set. Second, evaluation is often narrow and one-dimensional, relying either on user perception, surface-level NLP metrics, or domain-specific accuracy, but rarely integrating these perspectives into a unified assessment. Third, and most critically, few studies systematically assess educational generation using both semantic quality metrics and RAG-specific reliability indicators [12], such as faithfulness and hallucination rate.

2.4 Research Gaps

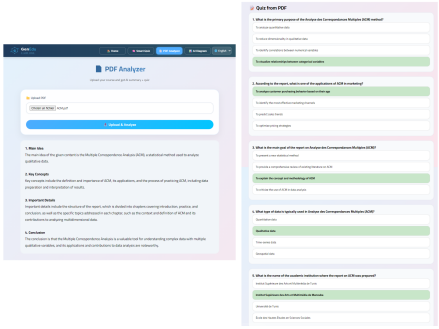
Three key research gaps motivate the development of GenEdu. First, existing approaches exhibit a limited integration between multi-functional educational support and retrieval grounding: systems typically either provide a single RAG-enhanced function with strong factual reliability but narrow pedagogical scope, or multiple learning functions without retrieval grounding, resulting in broader functionality but reduced consistency and reliability. GenEdu addresses this limitation by embedding RAG at the architectural core while supporting four pedagogically relevant functions within a unified framework. Second, evaluation practices in educational AI remain fragmented, often focusing exclusively on either user perception (e.g., surveys and usability measures) or automatic performance metrics (e.g., BLEU scores and classification accuracy), with little effort to bridge both perspectives. To overcome this limitation, GenEdu adopts a dual-layer evaluation strategy that combines user-centered assessment of perceived usefulness with automatic semantic evaluation, including RAG-specific measures, such as faithfulness and hallucination rate. Finally, although RAG is widely recognized as beneficial, there remains a lack of ablation-based evidence within educational contexts demonstrating its concrete impact on pedagogical reliability. GenEdu directly addresses this gap by comparing system performance with and without retrieval grounding, thereby quantifying its effect on semantic alignment, faithfulness, and hallucination reduction.



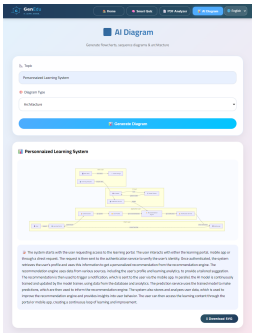
(a) Smart quiz module.



(b) AI-generated explanation module.



(c) PDF analyzer module.



(d) Diagram generator module.

Figure 1: GenEdu web interface: the four main learning modules.

3 System architecture and methodology

3.1 Framework overview

GenEdu is designed as a Web-based intelligent educational framework built around two complementary processing pipelines (See Figure 2). The first pipeline handles adaptive content generation (quizzes, explanations, diagrams), where outputs are customized to learner proficiency and performance. The second pipeline handles retrieval-augmented analysis, where all outputs are grounded in user-provided course materials (typically PDF documents), ensuring fidelity to instructor-provided content.

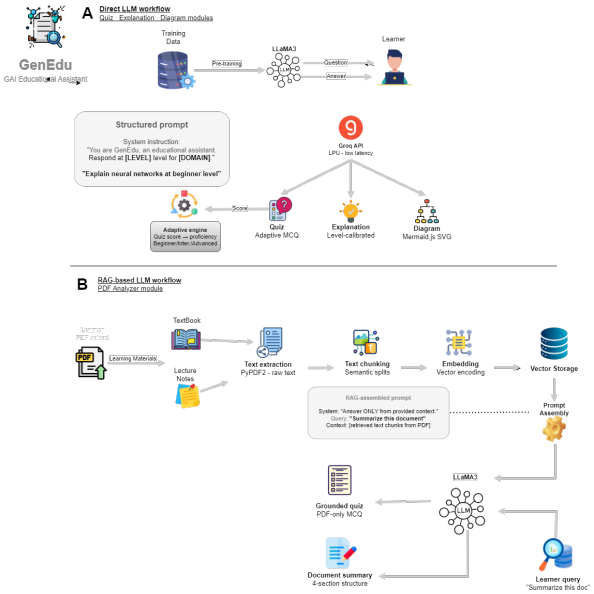


Figure 2: Overview of GenEdu framework architecture.

3.2 GenEdu modules and capabilities

GenEdu's Web interface (See Figure 1) comprises four main modules, each targeting a distinct pedagogical function:

- The *Smart Quiz Generator module* focuses on dynamic assessment through the generation of multiple-choice questions. Given a selected topic and a specified difficulty level (novice, intermediate, or advanced), the system uses a language model to produce context-appropriate quizzes and answer options. Learners receive immediate feedback on their responses, and performance outcomes are used to adapt subsequent difficulty levels, thereby supporting formative assessment and progressive skill development.

- The *AI-Generated Explanation Generator module* provides personalized conceptual explanations adapted to learner proficiency. Based on a selected topic and an estimated or user-defined proficiency level, relevant course passages are first retrieved using RAG. The language model, then, produces explanations grounded in these retrieved materials, adjusting linguistic complexity accordingly. This module primarily serves an instructional role by delivering on-demand and level-sensitive learning support.

- The *PDF Analyzer module* enables interaction with course documents through retrieval-augmented processing. After a PDF is uploaded, the system segments and indexes its content, allowing relevant sections to be retrieved in response to user queries. The language model, then, generates summaries, analytical insights, or quiz questions strictly grounded in the document content. This module supports reference and review activities, facilitating deeper engagement with instructional materials and self-assessment.

- The *Diagram Generator module* provides various learning support through the automatic creation of visual representations, such as flowcharts, sequence diagrams, architecture diagrams, and state machines. Given a conceptual description and a specified diagram type, the system produces a structured diagram definition that is subsequently rendered visually. This module enhances understanding of abstract or complex relationships by leveraging spatial and visual reasoning in the learning process.

3.3 Evaluation protocol: dual-layer approach

To assess both the pedagogical and technical value of GenEdu, we adopted a two-layer evaluation protocol that complements user perception with automatic reliability metrics.

Layer 1: User-centered evaluation and research questions The evaluation is driven by three main research questions.

- **RQ1**: How do users perceive the usability, content quality, and pedagogical usefulness of GenEdu across its four main functions, and do these perceptions differ depending on the function?
  To explore this, we conducted a user study with 106 participants recruited from academic and professional networks, including students, educators, and users with different levels of experience with AI tools. Participants answered a questionnaire made of 12 Likert-scale items covering five aspects: (i) usability and interface, (ii) quiz and content quality, (iii) PDF analyzer performance, (iv) AI diagram generation, and () personalization. We also checked the reliability of the instrument using Cronbach's alpha [13].

- **RQ2**: Does prior familiarity with AI tools influence how users evaluate GenEdu, and are there meaningful differences between AI-Familiar and AI-Occasional users?
  To answer this, we split participants into two groups based on their AI experience (AI-Familiar, $n = 58$; AI-Occasional, $n = 48$) and compared their responses using independent-samples t-tests.

- **RQ3**: Does retrieval grounding (RAG) actually improve the reliability of generated educational content, and how does it affect semantic alignment, factual consistency, and hallucination reduction?
  We addressed this by comparing a RAG-based configuration with a baseline LLM-only setup, focusing on how each approach performs in terms of content reliability and factual grounding.

Layer 2: Automatic semantic evaluation Automatic evaluation assessed the quality and reliability of generated outputs using both standard NLP metrics and RAG-specific reliability indicators. The evaluation benchmark consisted of 30 prompts covering the four main functionalities of the system. Reference outputs were derived directly from course materials.

- Standard NLP metrics included BLEU-2, which measures lexical overlap between generated and reference texts on a scale ranging from 0 to 1. ROUGE-1 and ROUGE-L F1 scores were also employed to assess the recall of overlapping n-grams and longest common subsequences, thereby capturing the degree of semantic preservation between generated and reference outputs. In addition, BERTScore F1 was utilized to evaluate semantic similarity through contextual embeddings, offering greater robustness to paraphrasing compared with traditional overlap-based metrics like BLEU and ROUGE.

- Two RAG-specific indicators were considered to further assess the reliability of the RAG process. *Faithfulness* measured the proportion of generated claims that were directly supported by the retrieved passages, with scores ranging from 0 to 1. A claim was considered faithful when a human annotator could identify a specific supporting passage within the retrieved material. Conversely, the *Hallucination Rate* quantified the proportion of generated claims that were either unsupported by the retrieved documents or contradicted the retrieved evidence, also on a scale from 0 to 1.

4 Results and analysis

4.1 Ablation study: impact of retrieval grounding

To isolate the effect of retrieval grounding, we compared system outputs generated with and without the RAG module. As shown in Table 1, including RAG substantially improved semantic quality and grounding reliability. BERTScore F1 increased from 0.7234 to 0.8965 (+23.9%). Faithfulness increased from 0.1500 to 0.6667 (+344.5%). The hallucination rate decreased from 0.8500 to 0.3333 (−60.8%).

Table 1: Ablation study: impact of RAG on response quality.

Metric	With RAG	Without RAG	Δ
BERTScore F1	0.8965	0.7234	+23.9%
Faithfulness	0.6667	0.1500	+344.5%
Hallucination Rate	0.3333	0.8500	−60.8%

These results confirm that retrieval grounding is not merely an auxiliary enhancement but a central design mechanism for improving the educational reliability of generated responses.

4.2 User-centered evaluation results

The user-centered study involved 106 participants. Results are presented in Table 2. The overall mean score was 4.2789/5.00, reflecting highly positive user perception. This strong satisfaction is consistent across all four functional dimensions, with the PDF Analyzer ($M = 4.34$) and Diagram Generation ($M = 4.32$) receiving the highest scores, while Personalization ($M = 4.24$) received the lowest (though still strong). This pattern suggests that users particularly appreciate the grounded, reference-oriented functions (document analysis, visual aids) while finding the adaptive aspects valuable but with slightly more room for refinement.

Besides, Cronbach's alpha yielded $\alpha = 0.952$, indicating excellent internal consistency [13]. This high alpha confirms that the five dimensions cohere around a single underlying construct (overall usability and pedagogical quality), justifying the use of a composite overall mean.

Table 2: User-centered evaluation results ($N = 106$).

Dimension	Mean	SD
Usability and Interface	4.2516	0.892
Quiz and Content Quality	4.2421	0.818
PDF Analyzer (RAG)	4.3443	0.860
AI Diagram Generation	4.3208	0.885
Personalization	4.2358	1.000
Overall Mean	**4.2789**	—

Additionally, the comparative analysis between user groups showed that AI-Familiar users rated GenEdu more positively across all dimensions, with statistically significant differences ($p < 0.05$). The largest gap was observed for the PDF Analyzer dimension, suggesting that prior exposure to AI tools helps users better exploit retrieval-grounded assistance.

4.3 Automatic semantic evaluation

Building on the ablation study, Table 3 presents the full automatic semantic evaluation across all 30 benchmark prompts under the RAG-enabled condition.

The system achieved a mean BERTScore F1 of 0.8965 ($SD = 0.0412$), ranging from 0.8640 to 0.9230, indicating strong semantic alignment across prompts. ROUGE-1 F1 reached 0.5373 and BLEU-2 reached 0.3266. Although lexical overlap metrics were moderate, BERTScore confirms high semantic consistency. The mean faithfulness score was 0.6667 and the mean hallucination rate was 0.3333, indicating that the system was generally successful in grounding outputs in retrieved material, while still highlighting room for improvement.

Table 3: Automatic semantic evaluation results ($n = 30$).

Metric	Mean	SD	Range
BLEU-2	0.3266	0.0892	0.1933–0.5244
ROUGE-1 F1	0.5373	0.1421	0.2462–0.6667
ROUGE-L F1	0.2985	0.1124	0.1569–0.4524
BERTScore F1	0.8965	0.0412	0.8640–0.9230
Faithfulness	0.6667	0.3333	0.0000–1.0000
Hallucination Rate	0.3333	0.3333	0.0000–1.0000

4.4 Discussion

Taken together, the two evaluation layers provide convergent evidence for the value of GenEdu. The user-centered study shows that the platform is perceived as

useful, clear, and educationally relevant across multiple learning support functions. The automatic evaluation confirms that these positive perceptions are accompanied by strong semantic performance and measurable gains in grounding quality. Most importantly, the ablation study demonstrates that RAG plays a decisive role in reducing hallucination and improving semantic alignment. These findings position GenEdu as a grounded pedagogical system in which personalization and reliability are jointly supported through architectural design, rather than a general-purpose generative assistant adapted to education.

5 Conclusion

This paper presents GenEdu, a retrieval-augmented generative AI framework for personalized learning that integrates adaptive content generation with retrieval grounding and a dual-layer evaluation strategy. Its main contributions are fourfold. First, it introduces a unified Web-based framework combining four pedagogical functions: adaptive quizzes, level-aware explanations, document-grounded analysis, and diagram generation. Second, it demonstrates the impact of retrieval grounding, with an ablation study showing gains in semantic alignment (+23.9%) and faithfulness (+344.5%), and a reduction in hallucinations (60.8%). Third, it proposes a hybrid evaluation protocol combining user study results with automatic metrics (BERTScore F1 = 0.8965, Faithfulness = 0.6667). Finally, it shows that AI familiarity significantly affects user perception, with AI-Familiar users rating the system higher ($p = 0.005$). Future work will focus on expanding the semantic evaluation dataset, strengthening long-term learner modeling through persistent cross-session tracking, and validating the framework across multiple institutions and subject domains.

References

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

[2] M. H. Y. Binhammad, A. Othman, L. Abuljadayel, H. Al Mheiri, M. Alkaabi, and M. Almarri, "Investigating how generative AI can create personalized learning materials tailored to individual student needs," *Creative Education*, vol. 15, no. 7, pp. 1499–1523, 2024.

[3] J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang *et al.*, "When large language models meet personalization: Perspectives of challenges and opportunities," *World Wide Web*, vol. 27, no. 4, p. 42, 2024.

[4] I. Pesovski, R. Santos, R. Henriques, and V. Trajkovik, "Generative AI for customizable learning experiences," *Sustainability*, vol. 16, no. 7, p. 3034, 2024.

[5] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "ChatGPT for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.

[6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, J. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2023, arXiv preprint arXiv:2312.10997.

[7] C. Sharma, "Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers," 2025, arXiv preprint arXiv:2506.00054.

[8] Z. N. Khlaif, W. A. Alkouk, N. Salama, and B. Abu Eideh, "Redesigning assessments for AI-enhanced learning: A framework for educators in the generative AI era," *Education Sciences*, vol. 15, no. 2, p. 174, 2025.

[9] B. Ma, M. A. Z. Khan, T. Yang, A. Polyzou, and S. Konomi, "How good are large language models for course recommendation in MOOCs?" 2025, arXiv preprint arXiv:2504.08208.

[10] J. S. Jauhiainen and A. Garagorry Guerra, "Generative AI and education: dynamic personalization of pupils' school learning material with ChatGPT," in *Frontiers in Education*. Frontiers Media SA, 2024, vol. 9, p. 1288723.

[11] C. Dong, Y. Yuan, K. Chen, S. Cheng, and C. Wen, "How to build an adaptive AI tutor for any course using knowledge graph-enhanced retrieval-augmented generation (KG-RAG)," in *Proceedings of the 14th International Conference on Educational and Information Technology (ICEIT)*. IEEE, 2025, pp. 152–157.

[12] H. Khosravi, M. R. Shafie, M. Hajiabadi, A. S. Raihan, and I. Ahmed, "Chatbots and ChatGPT: A bibliometric analysis and systematic review of publications in Web of Science and Scopus databases," *International Journal of Data Mining, Modelling and Management*, vol. 16, no. 2, pp. 113–147, 2024.

[13] D. George and P. Mallery, *SPSS for Windows Step by Step: A Simple Guide and Reference*. Allyn & Bacon, 2003.