ADVANCING STUDENT LEARNING WITH RAG-ENHANCED LARGE LANGUAGE MODEL CHATBOTS

Nguyen Viet Ha[1], *Tran Tuan Vinh*[1*]

Abstract: *The integration of artificial intelligence (AI) in education has spurred advancements in learning tools. This study presents the development and evaluation of an intelligent educational chatbot system based on Large Language Models (LLMs) enhanced with Retrieval-Augmented Generation (RAG) techniques. The RAG mechanism allows the chatbot to retrieve precise academic information from reliable knowledge bases, surpassing traditional search methods to generate evidence-based, contextually relevant responses. Utilizing the Llama 3.2 model as its foundation, the system was piloted using question sets from the subject Teaching Methods in Informatics for High School Education, demonstrating its suitability for educational applications requiring high accuracy. The chatbot effectively supports self-directed learning by answering queries, generating quizzes, and offering dynamic, personalized assistance. This study contributes to the practical application of RAG-enhanced models in personalized education, showing potential to reduce instructional workload and foster learner autonomy.*

Keywords: *Chatbot, large language models, retrieval-augmented generation, learning support, AI in education*

1. INTRODUCTION

Artificial intelligence (AI) is rapidly transforming various sectors [1-4], with education being a prime beneficiary [5]. The continuous advancement of AI-powered tools, particularly chatbots leveraging Large Language Models (LLMs), offers unprecedented opportunities to enhance learning experiences. Traditional pedagogical approaches, such as textbooks and conventional online courses, often struggle to provide the interactive and personalized support necessary for optimal learning outcomes. In contrast, machine learning models, and especially LLMs, can offer deeper insights and personalized assistance, addressing student learning difficulties individually [6, 7].

The potential of chatbots in education is widely recognized. It is posited that chatbots can enhance educational quality through personalized learning, engaging interactive environments, and real-time competency assessments with timely support [8-10]. Chatbots serve as readily available learning assistants, facilitating information retrieval, knowledge dissemination, and improved comprehension. Effective chatbot design can lead to a more seamless learning experience, while teachers can leverage student inquiries

[1] Hanoi Pedagogical University 2

[*] Corresponding author: trantuanvinh@hpu2.edu.vn

to refine knowledge bases through automated question analysis and response generation. Notably, students often prefer chatbots over traditional search engines due to their ability to provide direct answers [11-14].

Further exploration of chatbot applications in education reveals diverse implementations. Fernoagă et al. [15] explored chatbots as personalized educational assistants, employing a microservices architecture to create an automated adaptive teaching system. However, this system's limitation in multiple-choice questions highlights the need for more versatile interaction capabilities. Shen [16] examined the broader integration of generative AI and LLMs in education, emphasizing the potential for personalized learning and teacher support, while also acknowledging challenges such as over-reliance on automated systems, content accuracy, and ethical concerns related to data privacy and bias. Ajani et al. [17] provided an overview of AI's role in enhancing higher education, identifying applications like personalized learning and adaptive platforms, while noting significant implementation challenges. Gan et al. [18] further highlighted the potential of LLMs to address persistent issues in education, such as student diversity and unequal resource distribution, through personalized learning and intelligent tutoring.

A critical aspect of LLM implementation is addressing their inherent limitations, particularly in generating precise answers to specific queries. Lewis et al. [19] identified Retrieval-Augmented Generation (RAG) as a promising approach to improve LLM performance by enhancing response accuracy in domain-specific contexts. This is further substantiated by Das, Rangan, et al. [20], who demonstrated the effectiveness of integrating LLMs with knowledge-based retrieval systems in the medical domain. These findings underscore the potential of combining LLMs with structured knowledge bases to generate precise answers in specialized fields.

Despite the growing integration of LLMs in education, the widespread adoption of open-source LLMs remains limited. This research aims to address this gap by developing a chatbot based on an open-source LLM, integrated with RAG techniques. By combining the generative capabilities of LLMs with the precision of information retrieval, this chatbot aims to serve as an intelligent and adaptable learning companion, supporting students across diverse educational settings.

While challenges persist, the advancements in AI and LLM-based chatbots, particularly when coupled with RAG techniques, present substantial opportunities for educational transformation. By generating more precise and contextually relevant responses, these chatbots can significantly enhance the effectiveness of AI-driven learning support systems.

This study aims to:

- Develop an LLM-based chatbot with robust learning support capabilities.
- Implement RAG techniques to enhance response accuracy and relevance.
- Evaluate chatbot performance through quantitative metrics.

2. RESEARCH CONTENT

2.1. Data and Methods

2.1.1. Chatbot System Architecture

The chatbot workflow involves user queries being processed by an LLM integrated with a knowledge base stored in ChromaDB. The system retrieves relevant context before generating responses. When a user submits a query, the data is transmitted to the server, where it undergoes processing before being delivered back to the user's terminal. All system and user data are stored in a database for research and system enhancement purposes.

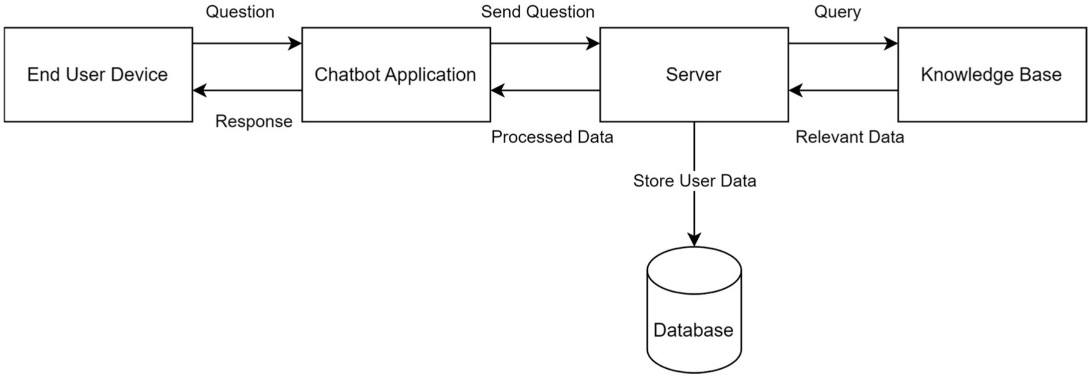

Figure 1. *Chatbot workflow diagram*

2.1.2. Retrieval-Augmented Generation (RAG)

RAG combines retrieval-based knowledge retrieval with LLM-based response generation. This approach ensures more accurate and contextually relevant responses. This technique integrates knowledge retrieval methods with vectorized data and large language models to generate accurate and contextually relevant responses for users. The system is implemented using ChromaDB's vector retrieval engine. The diagram below illustrates the RAG (Retrieval-Augmented Generation) process when a user queries specific topics.

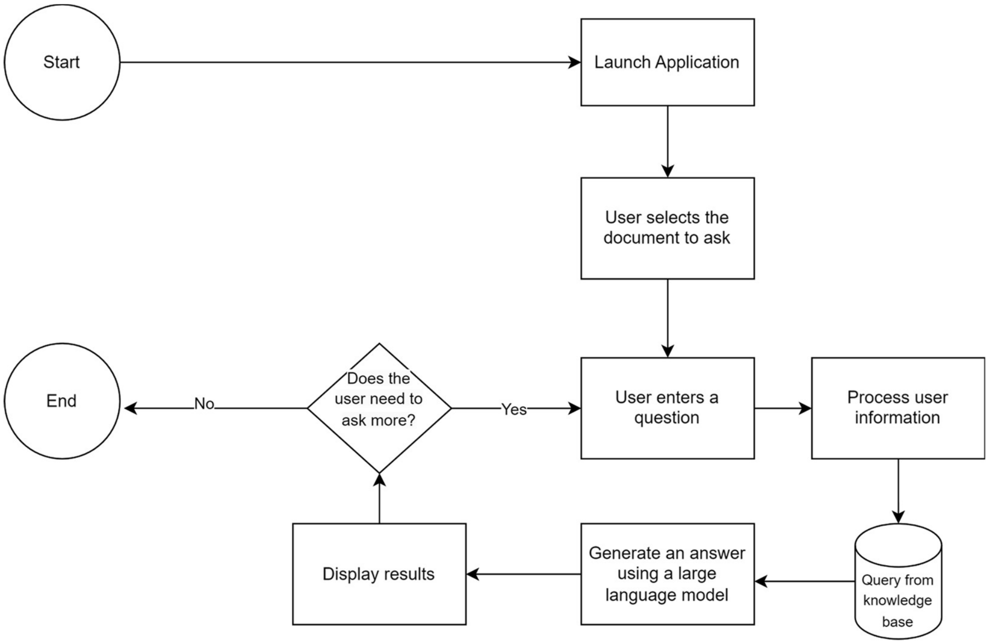
Start
Launch Application
User selects the document to ask
End
No
Does the user need to ask more?
Yes
User enters a question
Process user information
Display results
Generate an answer using a large language model
Query from knowledge base

Figure 2. *User interaction progress with the RAG system*

The chatbot's knowledge base is created by:

- Extracting content from PDF documents using PyPDF2.
- Splitting text into smaller segments with LangChain.
- Generating embeddings using the BAAI/bge-m3 model.
- Storing processed data in the ChromaDB vector database.

2.1.3. Evaluation

The evaluation of RAG systems demands methodologies that account for both retrieval and generative performance. DeepEval provides a robust and scalable framework for assessing these aspects, offering superior insights compared to traditional evaluation techniques. Its integration into RAG system assessment pipelines can lead to more reliable and effective AI-driven solutions. In this study, we use the three following parameters to evaluate the performance of the RAG system.

2.1.3.1. Answer Relevancy

Answer Relevancy measures the degree to which the generated response aligns with the user query. This parameter ensures that the chatbot effectively interprets user intent and produces responses that are contextually appropriate. High relevancy scores indicate that the system retrieves and utilizes pertinent information, improving user satisfaction and reducing misleading outputs.

2.1.3.2. Faithfulness

Faithfulness assesses whether the generated response remains factually accurate concerning the retrieved content. One of the primary challenges of generative AI is hallucination–where the model produces plausible but incorrect or unverifiable information. By incorporating this metric, we ensure that RAG systems generate reliable and fact-based responses, reinforcing trustworthiness in applications such as legal, medical, and financial domains.

2.1.3.3. Context Precision & Recall

The effectiveness of retrieval directly impacts the quality of generated responses. Context Precision & Recall quantify how well the retrieved documents contribute to the response. Precision evaluates whether retrieved documents are relevant to the query, while Recall measures whether all necessary documents have been retrieved. Balancing both ensures that the chatbot has access to comprehensive and accurate information, minimizing irrelevant content and improving overall response quality.

2.2. Results and Discussion

Using Llama3.2:3B and BAAI/bge-m3 embedding, chatbot responses were evaluated on 50 random questions on the topic of teaching methods for Informatics at the primary education level. The content of these questions was developed in alignment with the current General Education Curriculum issued by the Ministry of Education and Training of Vietnam in 2018. The primary source materials for question development were professional training documents and thematic materials on Informatics teaching methods for primary school teachers. The questions were divided into two categories: 25 questions requiring the ability to synthesize information from the documents, and 25 questions designed to assess the chatbot's reasoning capabilities.

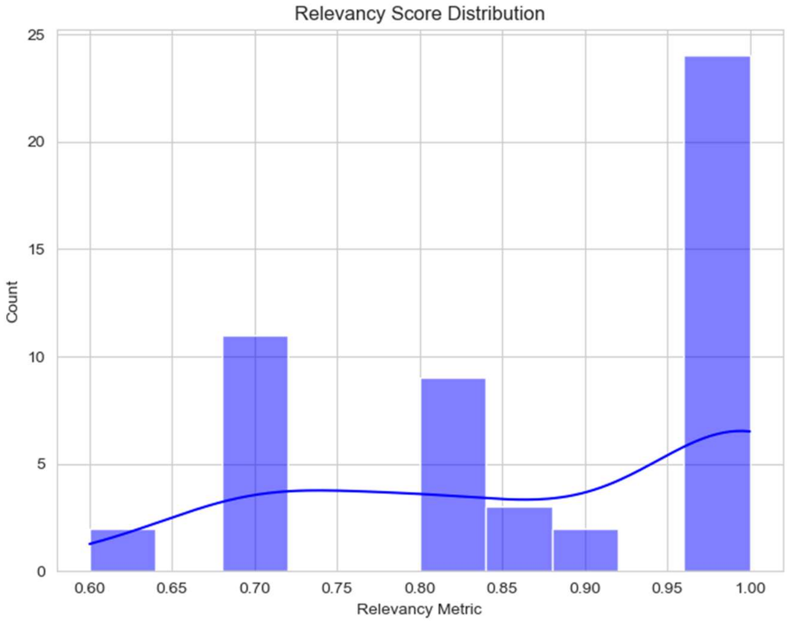

Figure 3. *Statistics on the relevancy metric of responses to 50 random questions on the topic of teaching methods for Informatics when using a chatbot powered by the Llama 3.2 3B model combined with the RAG method, utilizing the BAAI/bge-m3 embedding model.*

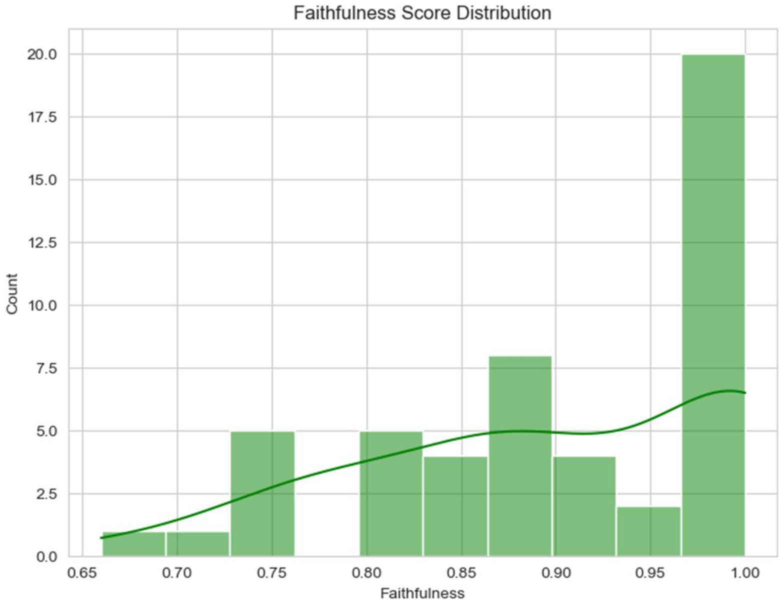

Figure 4. *Statistics on the faithfulness metric of responses to 50 random questions on the topic of teaching methods for Informatics when using a chatbot powered by the Llama 3.2 3B model combined with the RAG method, utilizing the BAAI/bge-m3 embedding model.*

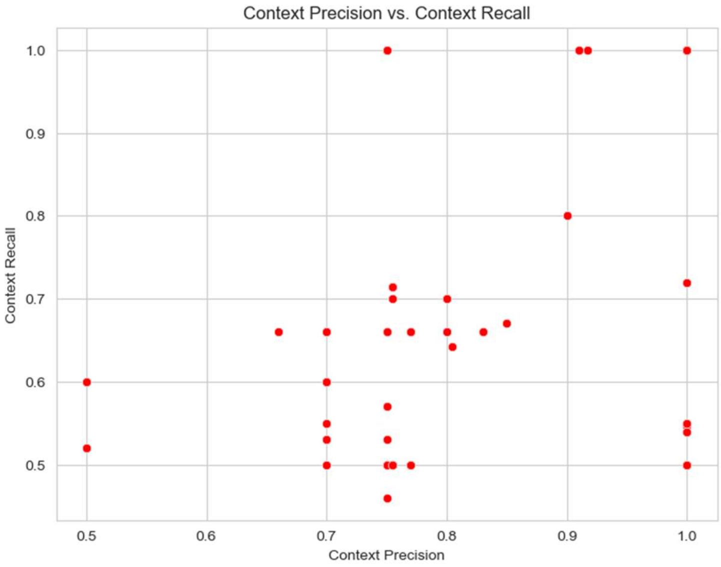

Figure 5. *Statistics on context precision and context recall of responses to 50 random questions on the topic of teaching methods for Informatics when using a chatbot powered by the Llama 3.2:3B model combined with the RAG method, utilizing the BAAI/bge-m3 embedding model.*

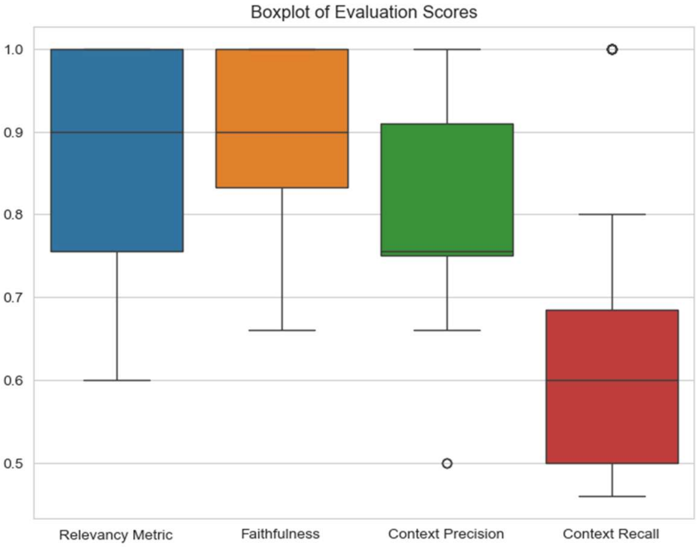

Figure 6: *Boxplot on 4 metrics of responses to 50 random questions on the topic of teaching methods for Informatics when using a chatbot powered by the Llama 3.2 3B model combined with the RAG method, utilizing the BAAI/bge-m3 embedding model.*

The dataset used for evaluation comprises 50 randomly selected questions related to teaching methods for informatics in secondary schools, with all questions formulated in

Vietnamese. Although LLaMA 3.2 does not natively support Vietnamese, the results remain acceptable.

Figures 3 and 4 highlight that the Answer Relevancy and Faithfulness scores of LLaMA 3.2: 3B, when combined with the Retrieval-Augmented Generation (RAG) approach, effectively address the challenges identified in prior studies regarding model accuracy in specialized domains. This suggests that the model can provide learners with a more personalized learning experience.

Figure 5 presents the scatter plot of Context Precision and Context Recall. Overall, the system demonstrates the ability to retrieve the necessary information for generating responses accurately. However, in some instances, the retrieved information remains insufficient.

Figure 6 illustrates the overall performance of the LLaMA 3.2: 3B model in responding to user queries. The evaluation results based on four key metrics—Answer Relevancy, Faithfulness, Context Precision, and Context Recall—demonstrate that LLaMA 3.2: 3B is fully capable of analyzing, synthesizing, and providing accurate responses to users. Moreover, from a cost-efficiency perspective, the model's significantly smaller parameter count compared to other models and its open-source nature make it a viable solution to challenges discussed in prior studies, particularly concerning response accuracy, efficiency, and privacy issues associated with closed-source models.

However, when using the DeepEval framework, the evaluation results rely on another LLM, which may sometimes necessitate human review, as the LLM may not fully capture all nuances of the transmitted information.

In summary, the effectiveness of the RAG approach in this system is acceptable for applications requiring high accuracy. Compared to the system proposed in [20], which integrates RAG with Hybrid Search and employs a closed-source model, our study demonstrates that a purely RAG-based approach combined with Vector Similarity Search and an open-source language model can yield competitive results. Specifically, using DeepEval benchmarking, the model in [20]—which employs Claude 3 Opus—achieves an Answer Relevancy score of 0.9, Faithfulness of 1, and Context Precision of 0.37. In contrast, our system achieves an average Answer Relevancy score of 0.9, Faithfulness of 0.9, and Context Precision of 0.8. These results underscore the effectiveness of the open-source LLaMA 3.2 3B model and the RAG approach with Vector Similarity Search for high-accuracy applications.

3. CONCLUSION

In conclusion, this research successfully developed and evaluated a chatbot employing open-source LLMs and a vector database, demonstrating its potential to support student learning through question answering. The Llama 3.2 3B model proved effective in a Retrieval Augmented Generation (RAG) system, providing accurate responses to common student inquiries. However, the study identified limitations in the RAG system's reliability, likely stemming from challenges in query vectorization. Optimizing query processing algorithms is crucial for future improvements. Furthermore, while the Llama 3.2 3B model performed well on routine questions, its performance on complex, reasoning-intensive queries was moderate, suggesting a need for further model refinement or alternative architectures to enhance performance in these areas. These findings highlight both the promise and the challenges of utilizing LLMs within RAG systems for educational support, pointing towards specific avenues for future research and development.

REFERENCES

1. D. Bit, S. Biswas, and M. Nag, *The impact of artificial intelligence in educational system*, Indo Am. J. Pharm. Res, vol. 11, pp.419-427, 2024, doi: 10.32628/IJSRST2411424.
2. T. K. Chiu, *The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney*, Interactive Learning Environments, vol. 32, no. 10, pp.6187-6203, 2024, doi: 10.1080/10494820.2023.2253861.
3. S. Rahayu, *The impact of artificial intelligence on education: Opportunities and challenges*, Jurnal Educatio FKIP UNMA, vol. 9, no. 4, pp.2132-2140, 2023, doi: 10.31949/educatio.v9i4.6110.
4. F. Kamalov, D. Santandreu Calonge, and I. Gurrib, *New era of artificial intelligence in education: Towards a sustainable multifaceted revolution*, Sustainability, vol. 15, no. 16, p. 12451, 2023, doi: 10.3390/su151612451.
5. X. Zhai et al., *A Review of Artificial Intelligence (AI) in Education from 2010 to 2020*, Complexity, vol. 2021, no. 8812542, pp.1-18, Apr. 2021, doi: 10.1155/2021/8812542.
6. A. Ahmed, S. Aziz, Alaa Abd-alrazaq, Rawan AlSaad, and J. Sheikh, *Leveraging LLMs and wearables to provide personalized recommendations for enhancing student well-being and academic performance through a proof of concept*, Scientific Reports, vol. 15, no. 1, Feb. 2025, doi: 10.1038/s41598-025-89386-2002E

7. S. Steinert, K. E. Avila, S. Ruzika, J. Kuhn, and S. Küchemann, *Harnessing large language models to develop research-based learning assistants for formative feedback*," Smart Learning Environments, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40561-024-00354-1.

8. İ. Əşrəfova, *Education and Chatbots: New Opportunities for Teachers and Students*, Journal of Azerbaijan Language and Education Studies, vol. 2, no. 2, pp. 39-49, 2025, doi: 10.69760/jales.2025001010.

9. Kayembe C. and Nel D., *Challenges and opportunities for education in the Fourth Industrial Revolution*," African Journal of Public Affairs, vol. 11, no. 3, pp.79-94, Sep. 2019, doi: 10.10520/EJC-19605d342e.

10. R. Winkler and M. Soellner, *Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis*, Academy of Management Proceedings, vol. 2018, no. 1, p. 15903, Aug. 2018, doi: 10.5465/ambpp.2018.15903abstract.

11. H. B. Essel, D. Vlachopoulos, A. Tachie-Menson, E. E. Johnson, and P. K. Baah, *The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education*, International Journal of Educational Technology in Higher Education, vol. 19, no. 1, Nov. 2022, doi: 10.1186/s41239-022-00362-6.

12. A. A. Mashinchi, *Chatbots as Classroom Assistants: A Qualitative Study on Teachers' Perspectives*, AI and Tech in Behavioral and Social Sciences, vol. 2, no. 2, pp. 12–19, 2024, doi: 10.61838/kman.aitech.2.2.3.

13. M. De, M. M. Chan, A. Garcia-Cabot, E. Garcia-Lopez, and Héctor Amado-Salvatierra, *The impact of a chatbot working as an assistant in a course for supporting student learning and engagement*, Computer applications in engineering education, May 2024, doi: 10.1002/cae.22750.

14. A. Silvervarg, C. Kirkegaard, Jens Nirme, M. Haake, and Agneta Gulz, *Steps towards a Challenging Teachable Agent*, Lecture notes in computer science, pp. 410–419, Jan. 2014, doi: 10.1007/978-3-319-09767-1_52.

15. Paul vlad Fernoaga, Cristinel Gavrila, F. Sandu, and Georgealex Stelea, "Intelligent education assistant powered by chatbots," Apr. 2018, doi: 10.12753/2066-026x-18-122.

16. S. Shen, *Application of large language models in the field of education*, Theoretical and Natural Science, vol. 34, no. 1, pp.147-154, Apr. 2024, doi: 10.54254/2753-8818/34/20241163.

17. O. A. Ajani, B. Gamede, and T. C. Matiyenga, *Leveraging artificial intelligence to enhance teaching and learning in higher education: Promoting quality education*

and critical engagement, Journal of Pedagogical Sociology and Psychology, Oct. 2024, doi: 10.33902/jpsp.202528400.

18. W. Gan, Z. Qi, J. Wu, and Jerry Chun-Wei Lin, *Large Language Models in Education: Vision and Opportunities*, Dec. 2023, doi: 10.1109/bigdata59044.2023.10386291.
19. P. Lewis *et al.*, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, arXiv.org, Apr. 12, 2021. https://arxiv.org/abs/2005.11401
20. R. Das *et al.*, *Improved precision oncology question-answering using agentic LLM*, medRxiv (Cold Spring Harbor Laboratory), Sep. 2024, doi: 10.1101/2024.09.20.24314076.

NÂNG CAO KHẢ NĂNG HỌC TẬP CỦA SINH VIÊN VỚI CHATBOT SỬ DỤNG MÔ HÌNH NGÔN NGỮ LỚN KẾT HỢP KỸ THUẬT RAG

Nguyễn Việt Hà, Trần Tuấn Vinh

Tóm tắt: *Việc tích hợp trí tuệ nhân tạo (AI) vào lĩnh vực giáo dục đã góp phần thúc đẩy sự tiến bộ trong việc phát triển các công cụ hỗ trợ học tập. Nghiên cứu này trình bày quá trình phát triển và đánh giá một hệ thống chatbot giáo dục thông minh dựa trên Mô hình ngôn ngữ lớn (LLM) kết hợp kỹ thuật Tạo sinh dựa trên truy xuất tăng cường (Retrieval-Augmented Generation – RAG). Kỹ thuật RAG cho phép chatbot truy xuất thông tin học thuật chính xác từ các cơ sở dữ liệu chứa kiến thức đáng tin cậy, vượt qua các phương pháp tìm kiếm truyền thống để tạo ra các phản hồi có căn cứ và phù hợp với ngữ cảnh. Hệ thống ứng dụng mô hình Llama 3.2, được thử nghiệm với bộ câu hỏi của môn Phương pháp dạy học môn Tin học ở trường phổ thông và kết quả cho thấy hệ thống có thể đáp ứng được các yêu cầu đòi hỏi độ chính xác cao. Chatbot có thể hỗ trợ hiệu quả việc tự học bằng cách trả lời các truy vấn tạo câu hỏi và cung cấp hỗ trợ thích ứng với nhu cầu của người học. Nghiên cứu này đóng góp vào việc ứng dụng thực tiễn các mô hình kết hợp RAG trong giáo dục nhằm giảm tải khối lượng giảng dạy và nâng cao tính tự chủ cho người học.*

Từ khóa: *Chatbot, mô hình ngôn ngữ lớn, tạo sinh dựa trên truy xuất tăng cường, mỗ trợ học tập, AI trong giáo dục.*

(Ngày Tòa soạn nhận được bài: 10-3-2025; ngày phản biện đánh giá: 21-3-2025; ngày chấp nhận đăng: 16-4-2025)