HNUE JOURNAL OF SCIENCE
Natural Sciences 2025, Volume 70, Issue 4, pp. 43-56
This paper is available online at http://hnuejs.edu.vn/ns
DOI: 10.18173/2354-1059.2025-0053

BUILDING A VIETNAMESE MATH CHATBOT BASED ON RAG AND LLM: SYSTEM DESIGN, IMPLEMENTATION AND EXPERIMENTAL EVALUATION

Pham Van Khanh[1] and Pham Vu Anh Tuan[2,*]

[1]*Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi city, Vietnam*

[2]*OPT Student of the Department of Computer Science, DePauw University, Indiana State, The USA*

[*]Corresponding author: Pham Vu Anh Tuan, e-mail: tuanpva.2001@gmail.com

Received November 6, 2025. Revised December 18, 2025. Accepted December 30, 2025.

Abstract. In recent years, large language models (LLM) and Retrieval-Augmented Generation (RAG) techniques have opened up new opportunities for the development of intelligent learning assistant systems. Nevertheless, the direct application of LLMs to Vietnamese Mathematics still has several limitations, including the illusion effect, a lack of knowledge base, failures to adhere to the Vietnamese education curricula, and difficulty in proceessing complex-related image issues. This paper presents the design, implementation, and experimental evaluation of a Vietnamese Mathematics Chatbot system based on the RAG architecture combined with LLM. The system comprises: (i) a pineline for collecting and standardizing Mathematics data from textbooks, exam questions and reference materials; (ii) a Milvus vector database to store embeddings generated by the BGE-m3 model; (iii) a multi-task pipeline coordinated by LangGraph; (iv) the inference component uses the Qwen3-VL-8B model implemented via vLLM; and (v) the WebUI user interface supports multimodal queries (text + image). Experimental results show that the system competently delivers detailed solutions, maintains the conversation flow, and significantly mitigates hallucination compared to pure LLM. The system also demonstrates potential for application in teaching and learning Mathematics, especially in situations requiring accurate knowledge retrieval and step-by-step explanations. These results suggest directions towards developing specialized learning assistants within the context of Vietnamese education.

Keywords: Math Chatbot, generative artificial intelligence, RAG, large language models, Milvus, vLLM, LangGraph, mathematics education.

1. Introduction

Mathematics plays a central role in the Vietnamese general education system, not only as a compulsory subject throughout all levels but also as the foundation for major exams such as transfer exams, high school graduation exams, and university entrance exams. In accordance with the current educational reform orientation, the requirements to enhance critical thinking, problem-solving skills, and self-learning capacity pose significant challenges for both teachers and students. In that context, artificial intelligence technologies, primarily large language models, are opening up new opportunities to support personalized learning and optimize the teaching-learning process [1], [2].

The rapid development of generative AI models such as GPT, LLaMA, Qwen, and Mistral allows chatbots not only to answer questions but also explain, propose methods, simulate conversations, and interact in multiple ways. Numerous studies have demonstrated the effectiveness of chatbots in education, including supporting learning, providing instant feedback, and enhancing the engagement of learners [3], [4]. In the field of Mathematics, the use of chatbots can assist students in approaching solutions step by step, understand solution methods, and diminish dependence on tutors or traditional materials [5].

However, the direct application of LLM to Vietnamese Mathematics reveals several critical limitations. Firstly, the phenomenon of "hallucination" – where the model confidently generates incorrect solutions – causes difficulties in educational environments where accuracy is a key factor [6]. Secondly, multilingual LLM training models often do not closely follow the Vietnamese Mathematics program, leading to non-standard answers or the use of inappropriate terminology. Thirdly, persistent limitations remain in the ability to understand and process images containing mathematical texts, geometric drawings, or complex formulas of general models [7]. Finally, the content generated by LLM often lacks a clear basis, which is difficult to verify, or is not linked to official knowledge sources such as textbooks or exam questions.

To address these limitations, the Retrieval-Augmented Generation architecture has emerged as an effective solution, combining the generative power of LLM with a tightly controlled knowledge base [8]. Instead of allowing the model make autonomous inferences from training data, RAG forces the model to retrieve relevant documents from a structured knowledge base (such as a vector database) before generating an answer. This minimizes artifacts, enhances accuracy, and enables the system to update its knowledge without retraining the pre-trained model. RAG is particularly appropriate for fields that demand high accuracy, such as medicine, law, and STEM education [9].

In the context of Vietnamese education, where the curriculum–textbooks–output standards are clearly defined, RAG provides a special advantage: chatbots can respond based on official documents, including textbooks, exam questions, exercises, topics, and reference materials. This establishes a transparent, well-founded learning assistant system that strictly aligns with the strategic orientation of the Ministry of Education and Training. Furthermore, the use of vector databases such as Milvus, combined with a modern embedding model (BGE-m3), enables the system to accurately retrieve similar problems,

related formulas, or sample solutions – something that LLMs alone struggle to execute effectively [10].

A further critical challenge in the Math chatbot system is interpreting multimodal inputs. Vietnamese students often take photos of problems from textbooks, test papers, or scratch paper, which may contain LaTeX formulas, special symbols, or drawings. Therefore, a multimodal model like Qwen3-VL-8B is a suitable choice due to its ability to recognize the visual information and transform it into structured text, helping the chatbot handle the problem more completely and accurately [11]. Moreover, LangGraph is utilized for pipeline coordination to facilitate a robust and scalable system framework. LangGraph models the query-handling process as a state graph, enabling the management of multiple steps such as query analysis, knowledge retrieval, prompt construction, LLM inference, and response by each independent node. This makes the system easy to maintain, reutilize, and extend in the agent-based framework - the dominant trend of contemporary AI [12]. Finally, deploying the system on Docker Compose ensures a synchronized, reproducible environment and simplifies the testing process. With the above motivations, this paper proposes the design, implementation, and experimental evaluation of a Vietnamese Mathematics chatbot based on RAG and LLM. The main contributions of the research include: (i) Building RAG pipeline in a multi-layer architecture, that supports multi-modal queries, and maintains conversational context; (ii) Building a Vietnamese Mathematics knowledge base in JSONL format to be converted into a Milvus vector database; (iii) Integrating the Qwen3-VL-8B multimodal model via vLLM to process text and problem images simultaneously; (iv) Experimentally evaluating accuracy, halucination mitigation, solution coherence, and multi-turn conversation quality; (v) Proposing a set of criteria for evaluating Math chatbots in the context of Vietnamese education.

This research contributes to laying the foundation for AI learning assistant systems that strictly align with the Vietnamese curriculum, are safe, accurate, and verifiability The RAG-based Math Chatbot not only assists students in grasping solution methodologies but can also serve as a useful support tool for teachers in lesson planning and grading. Taken together, this research establishes a trajectory towards the development of specialized educational AI systems tailored to the national context and meeting the requirements of pedagogical innovation in the digital age.

2. Content

2.1. Related work

Research on educational chatbots, knowledge retrieval-based question-answering systems, and large language models has proliferated in recent years. This section presents an overview of prominent research areas, including: (i) chatbots in educational settings; (ii) the application of large language models in learning Mathematics; (iii) Retrieval-Augmented Generation techniques and knowledge-based question-answering systems; and (iv) multimodel models and mathematical image processing. These studies form a critical foundation for the development of a Vietnamese Math Chatbot based on RAG and LLMs.

2.1.1. Chatbots in education

Educational chatbots have been widely deployed across various fields due to their ability to provide personalized support, deliver instant feedback, and alleviate teachers' workload. Several systematic reviews have demonstrated that chatbots strengthen students' learning motivation and engagement [3]. Huang et al. indicated that language chatbots facilitated their learning efficiency and mitigated anxiety when practicing foreign languages [4]. Other studies have focused on developing chatbots as "learning assistants" capable of explaining concepts, suggesting methodologies, and providing tailored, appropriate examples aligned with the learner's proficiency [3], [5].

Despite these benefits, traditional rule-based or retrieval-based chatbots are limited by their reliance on predefined datasets, resulting in a lack of flexibility and generalization. Consequently, the advent of large language models has revolutionized chatbot capabilities, enabling more natural conversations and deeper understanding of user intent [1], [2].

2.1.2. LLM application in supporting Math learning

In recent years, LLMs have been widely studied for automated problem-solving, including solution generation, task analysis, and step-by-step reasoning simulation. Numerous systems, such as MathGPT, GPT-4, and MATH-BERT, have demonstrated the capacity to solve problems ranging from basic to advanced levels [6] effectively. However, studies have also indicated that LLMs are prone to reasoning errors, especially when handling mathematical expressions with complex structures or those requiring multi-step reasoning [6], [7].

Models such as Qwen-VL [11] and PaLM-E have extended their capabilities to multimodal tasks, enabling the simultaneous recognition of text and formulas in images. This is particularly critical for Mathematics learners, as the problems are frequently sourced from textbooks or handwritten documents. However, these systems still exhibit limitations when applied to the Vietnamese educational environment due to the scarcity of Vietnamese-specific training data and a lack of alignment with the national Math curriculum.

2.1.3. Retrieval-based question-answering system and RAG architecture

Retrieval-Augmented Generation was introduced by Lewis et al. as an architecture that combines the generative capabilities of LLMs with external information retrieved from a structured knowledge base [8]. Subsequently, Karpukhin et al. developed Dense Passage Retrieval (DPR), which significantly enhances retrieval performance using semantic embeddings [9]. Modern embedding models such as BGE-m3 further advance cross-lingual and cross-domain retrieval capabilities [10], making RAG a pivotal framework for AI systems that require high accuracy and factual grounding.

Recent studies have demonstrated that RAG is particularly effective in domains that require high-fidelity knowledge: medicine, law, and STEM education [8], [9]. The information retrieval mechanism allows the model to generate fact-based answers, mitigating hallucinations, a common problem of LLMs [6]. For Mathematics education, RAG allows the system to leverage curated knowledge from textbooks, standardized exam questions, or academic documents, thereby ensuring factual relevance and alignment with the curriculum.

In the context of Vietnam, there is currently a scarcity of research applying RAG to develop specialized Math chatbots, especially Vietnamese chatbots. This represents a research gap that this study aims to address.

2.1.4. Knowledge management via vector database and Milvus

Vector database retrieval systems have been widely adopted to store embeddings and execute low-latency top-k similarity searches. Milvus, a leading open-source platform, provides a diverse set of indexing algorithms, dynamic schema support, and distributed scalability [10]. In existing literature, Milvus is frequently utilized as a core component of RAG architectures for document retrieval, question-answering, recommendation systems, and semantic search.

The integrationof Milvus into the Math assistant system enables the efficient retrieval of similar problems, mathematical formulas, and reference solutions from the knowledge base formatted in JSONL format. This constitutes a significant advantage over standard LLM chatbots that lack retrieval capabilities

2.1.5. Multimodal modeling and image processing

Qwen-VL [11], BLIP-2, PaLI-X, and several other multimodal models have extended the capability to process visual inputs containing text, symbols, and mathematical expressions. This enhances the ability of chatbot systems better support geometric problems, graphs, tables, and textbooks based exercie. Nevertheless, research in this area remains limited, particularly within the general education setting. In these environments, mathematical diagrams possess unique structures that require models to accurately interpret spatial components and geometric properties, such as points, segments, angles, and relationships, including parallelism and perpendicularity.

Research by Nguyen et al. [7] indicates that vision-language models are prone to erroneous conclusions when processing noisy images or those containing intricate details. Consequently, integrating multimodal LLMs with RAG represents a promising approach to enhancing the accuracy in processing image-centric mathematical problems.

2.1.6. Inference coordination system using LangGraph

LangGraph is a sophisticated framework that organizes LLM workflows as stateful graphs [12], facilitating the construction of complex pipelines with conditional logic, iterative loops, or parallel agentic operations. Within the RAG ecosystem, LangGraph is employed to coordinate tasks such as query classification, multi-modal data processing, knowledge retrieval, and LLM-based solution generation from LLM. Recent studies have demonstrated that LangGraph provides superior scalability, optimized letency improved speed, and easier maintenance compared to traditional linear pipelines.

2.2. Research methods

This section presents the methodological approach deployed to develop a Vietnamese Math Chatbot based on a large language model integrated with a Retrieval–Augmented Generation (RAG) architecture. The methodology is implemented around two primary workflows: the ingestion of and normalization of data to establish a structured mathematical knowledge base, and the query processing pipeline designed to generate answers that are accurate, coherent, and coherent answers aligned with the Vietnamese

curriculum. The theoretical components analyzed in the previous section are used as the foundation for the proposed architecture.

2.2.1. Overall architecture of the system

The Math Chatbot system is structured as a multi-layered pipeline (Figure 1), where each layer performs modular yet interconnected functions. The entire pipeline operates via a state management mechanism, which facilitates context processing maintain dialogue history, and enables decision-making at each execution step consistently. Operations are executed within a LangGraph-based state machineensuring architectural clarity, scalability, and data flow optimization.

The architecture consists of five functional layers: the data layer, the vectorization layer, the knowledge retrieval layer, the inference layer, and the interface layer. These layers operate sequentially from initial query input to final result output. User queries, comprising either text or image, are preprocessed and normalized before entering the pipeline. The system then categorizes the query and determines whether the problem requires knowledge extraction from the mathematical data warehouse. When a RAG-based search is triggered, the system retrieves relevant contextual fragments via MilvusDB before building an argumented prompt to send to the large language model. This process ensures that the model's inferences are grounded not only in pre-trained parameters but also in vetted, domain-specific data.

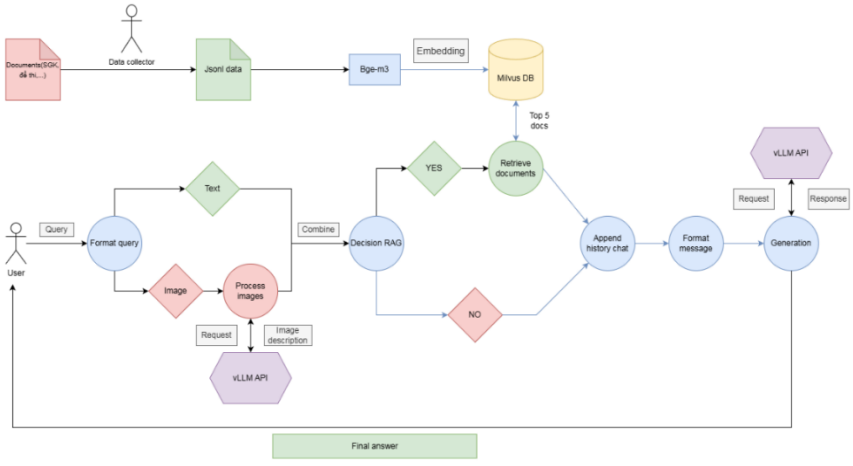

Figure 1. Overall architecture of the system

2.2.2. Building a Vietnamese mathematical knowledge base

To enable the Chatbot to provide responses aligned with Vietnamese pedagogical standards, the development of specialized knowledge is imperative. This knowledge base is formed through three distinct phases: document acquisition, JSONL structure standardization, and vectorization via a pre-trained embedding model.

The initial phase involved aggregating a large amount of mathematical materials from various sources, including standard textbooks, workbooks, national exams, mock exams, topical review papres and supplementary advanced materials. All sources are systematically categorized by topic, grade level, and problem typology to ensure structural consistency and retrieval efficiency.

To ensure data integrity, documents are preprocessed by removing redundantmetadata formatting mathematical expressions using LaTeX syntax. Each problem is subsequently structured into a discrete JSON entry comprising the problem statement and its corresponding detailed solution. This format facilitates seamless integration for both system fine-tuning and efficient retrieval.

For example:

{ "problem": "Given a,b,c>0 and a+b+c=1. Prove that $\\sum\\frac{a}{b+c}\\ge \\frac{3}{2}$" ,

"solution" : "Applying AM-GM: $b+c\\le \\frac{(b+c)+(a)}{2}=\\frac{1}{2}$ so $\\frac{a}{b+c}\\ge 2a$. Adding the three inequalities yields $\\sum\\frac{a}{b+c}\\ge 2(a+b+c)=2\\ge \\frac{3}{2}$. " }

In the final phase, the entirety of the JSONL dataset is transformed into high-dimensional semantic vectors using the BGE-m3 model. Each vector encodes the semantic representation of a problem or solution, which is subsequently indexed in Milvus to facilitate similarity-based retrieval. A highly-optimized search index is implemented to ensure that the sub-second query latency system and high precision, maintaining system performance even as the data repository scales.

2.2.3. User query processing and RAG pipeline

When a user submits a question, the system generates a new session state and receives all input data, including the query text, the question image (if any), and the most recent conversation history. The query is then passed through a preprocessing step to remove irrelevant characters, normalize the mathematical expressions, and separate the image and text components.

If the query contains an image, the Qwen3-VL-8B model is used to analyze the visual content. The model converts the image into standard text through character recognition, formula extraction, and geometric object representation. This descriptive text is merged with the user query to form a complete and comprehensive query.

A query classification module (Decision RAG node) is used to determine whether the query requires knowledge retrieval from the database. New problems, theoretical knowledge requests, and exercises requiring solutions trigger the RAG mechanism. Conversely, requests for dialogue, further explanation, or continuation of previous information may not require additional retrieval.

When a RAG is needed, the query is vectorized and sent to Milvus to search for the semantically closest data fragments. The five most similar documents are retrieved to help build the context for the model. This dataset serves as the background knowledge for the model prompt generation step.

The final prompt is generated by concatenating the system prompt, the knowledge context, the conversation history, and the current query. The prompt design prioritizes coherence, requiring the model to fully present the solution steps, use standard LaTeX, and avoid confusion between unrelated content. The prompt is then sent to the Qwen3-VL-8B model running on the vLLM framework to generate a solution.

The model returns a detailed solution including problem analysis, step-by-step reasoning, mathematical formulas, and a conclusion. The response is saved in the conversation history so that the system can continue to maintain subsequent exchanges.

2.2.4. Deploying the system in a containerized environment

A primary objective of this research was to build a system characterized by seamless deployment, reproducibility, and production readiness. Consequently, the entire architecture was deployed using the Docker Compose framework. This approach allows each service to be encapsulated within independent containers, enabling streamlined version control and scalability as needed.

The deployment environment consists of a standalone Milvus cluster, running etcd and MinIO for metadata coordination and object storage management. The inference cluster includes a vLLM serving large language models, a pipeline service that handles all LangGraph orchestration logic, and a WebUI for users. All services reside on the same local network, ensuring minimal latency and reliable connections between them.

Using Docker allows the system to launch quickly on many different machines without environmental or system-level dependency conflicts, meeting the requirements for testing, evaluation, and operational deployment.

2.3. Results and development directions

2.3.1. System implementation results

The implementation process shows that the Math Chatbot system based on RAG and LLM operates stably, meeting the requirements of accuracy, scalability, and user-friendliness. One of the most notable results is that the system supports multiple problem-solving models (Figure 2), each optimized for a specific class of problems, such as inequalities, equations - systems of equations, derivatives - differentials, or probability - statistics. Partitioning models by problem allows the Chatbot to utilize more appropriate reasoning strategies, thereby improving the accuracy and its ability to recognize the structural characteristics of the problem. Users, especially students and teachers, can easily select the model that aligns with their objectives, thereby minimizing errors and reinforcing the reliability of the solution.

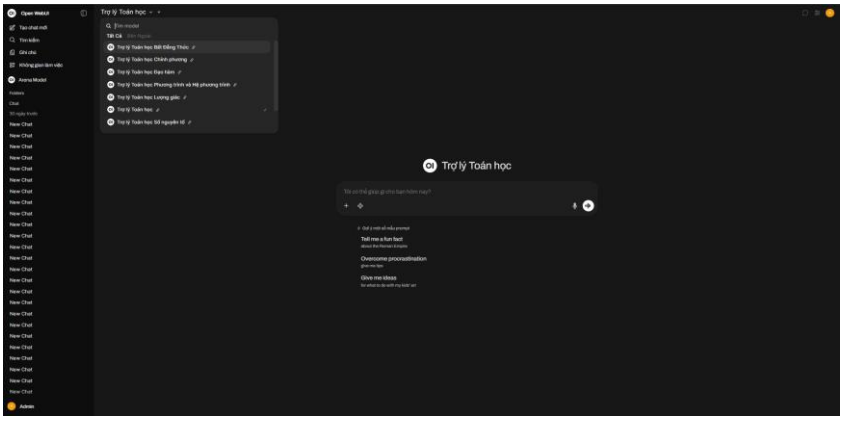

Figure 2. The system supports multiple models for each type of problem

Another notable result is that the system can provide comprehensive mathematical solutions (Figure 3), explained in clear, logical steps. Owing to the RAG pipeline combined with the Qwen3-VL-8B language model, the Chatbot not only answers the questions but also performs problem analysis, identifies the appropriate solution strategy, explains each calculation step, and renders formulas using standard LaTeX. The model's ability to generate coherent and accessible solutions plays a crucial role because this is one of the main needs of learners when using learning support systems. Experimental results show that the system's solutions are not only mathematically correct but also aligned with the familiar presentation style found in textbooks and exams in Vietnam.

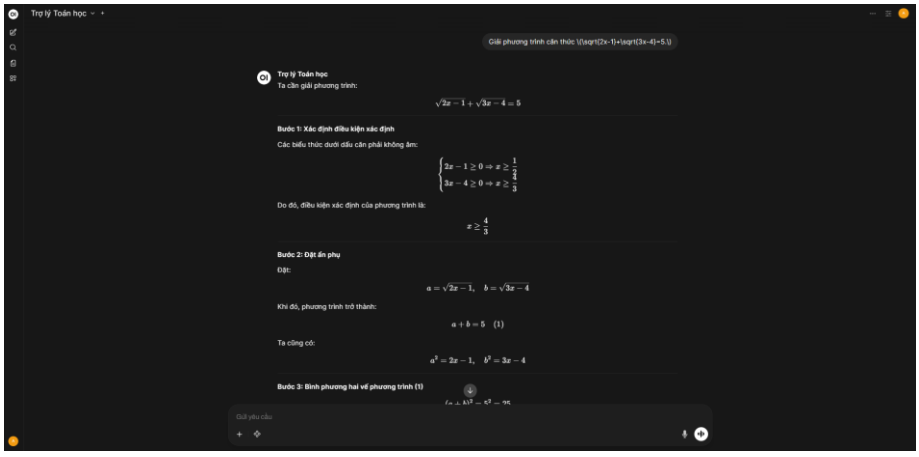

Figure 3. The system provides detailed solutions

The system also demonstrates the ability to process multimodal input data efficiently (Figure 4). When the user submits an image of the problem, the Qwen3-VL-8B vision module accurately recognizes textual content, mathematical expressions, and diagrammatic structures. The extracted text is integrated with the original query to construct a complete input, allowing the Chatbot to solve problems even from scanned or photographed textbook pages. This feature is particularly important in the context of high school students frequently using their phones to capture problem statements instead of entering them manually.

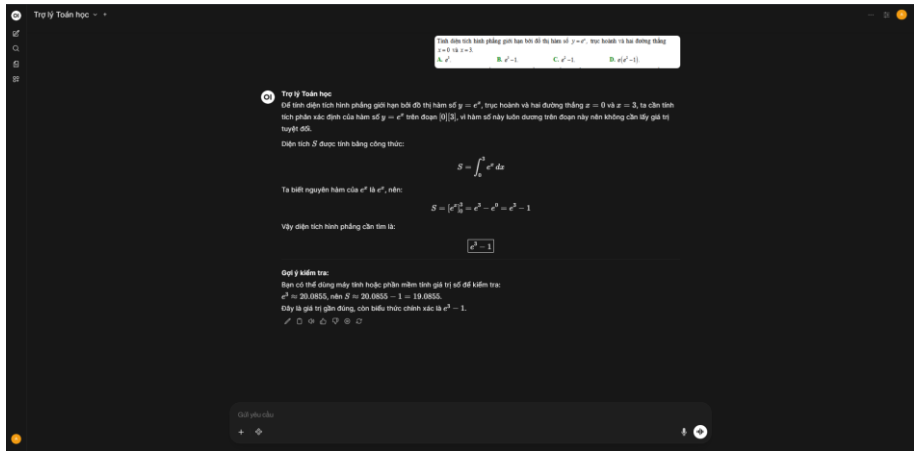

Figure 4. Multimodal input support system

Another significant strength is the system's ability to maintain long-term conversational context (Figure 5, Figure 6, Figure 7). The mechanism of storing and reprocessing conversation history helps Chatbot distinguish between sequential queries and standalone questions. Thanks to this context-awareness capability, the system avoids repeating information that has already been provided, and can deliver contextually appropriate responses in natural conversational situations, such as when the user asks to explain a step or expand on an existing solution. This contributes to an enhanced interactive experience, making the Chatbot more like a genuine learning assistant than a single-response tool.

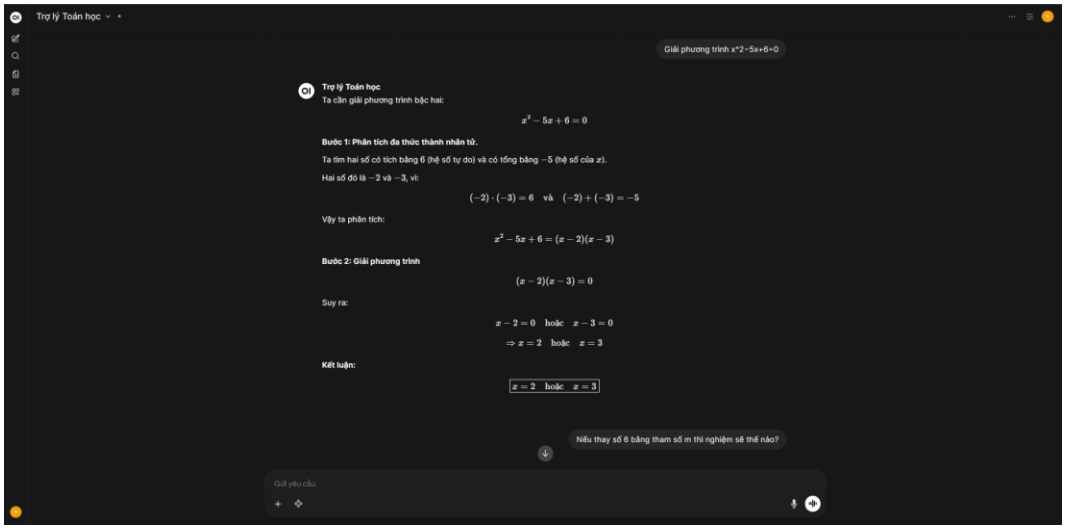

Figure 5. Multi-chat support system (1)

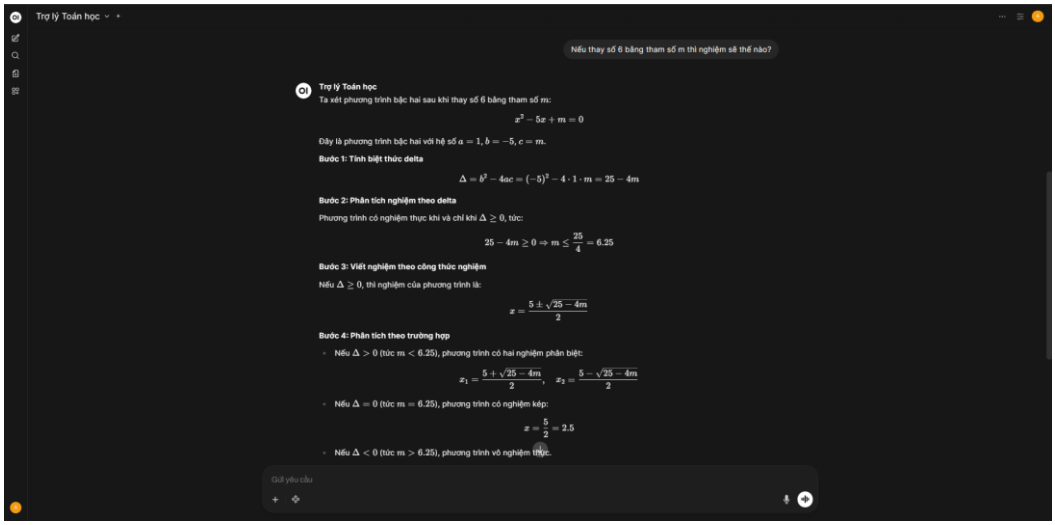

Figure 6. Multi-chat support system (2)

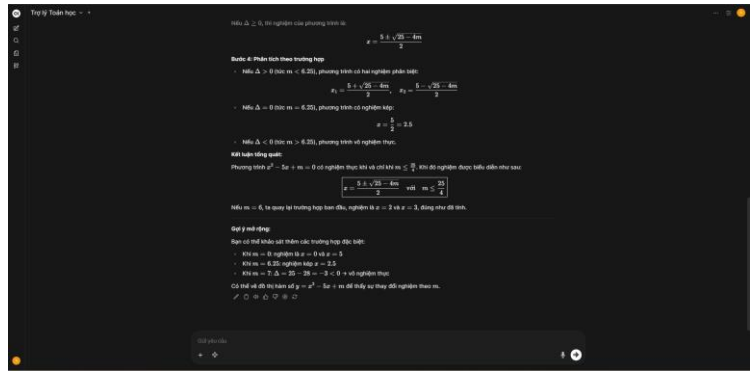

Figure 7. Multi-chat support system (3)

The implementation illustrates that the system has achieved the following main advantages: supporting multiple problem-solving models, multi-method processing, step-by-step solution generation, robust context preservation, and stable operation on containerized infrastructure. These results confirm that the RAG architecture, integrating LLMs with a vector database, is completely suitable for building a reliable Vietnamese math assistant.

2.4. Dataset & evaluation protocol

The evaluation dataset comprises 240 mathematics questions aligned with the national curriculum, designed to reflect realistic queries encountered by educational mathematics chatbots and to evaluate generalization performance. The questions span three educational levels: Primary (30%), Lower Secondary (40%), and Upper Secondary (30%), and cover core domains including Arithmetic, Algebra, Geometry, and Statistics–Probability. The dataset includes a balanced mix of multiple-choice (35%), short-answer (40%), and open-ended questions (25%), enabling evaluation of both final-answer accuracy and reasoning quality. Each question is paired with a ground-truth solution, specified either as a final result or as key solution steps, allowing for mathematically equivalent valid responses.

**** Experimental settings***

We evaluate the system under three configurations to isolate the contribution of the proposed approach: (i) a Baseline LLM that generates answers directly from pretrained knowledge, (ii) Baseline LLM + RAG, which intergates retrieved mathematical references as contextual input, and (iii) the Proposed Math Chatbot, which incorporates context selection, controlled step-by-step reasoning, and rule-based constraints to mitigate hallucination and enhance accuracy. All configurations are tested on the same question set using uniform generation settings, ensuring a fair and controlled comparison.

**** Evaluation metrics***

The mathematics chatbot is evaluated using four quantitative metrics: Exact Match (EM), quantifying the proportion of responses whose final answers exactly match the ground truth; Answer Accuracy, reporting the proportion of solutions classified as *Correct* on a three-level scale (Incorrect, Partially Correct, Correct); Hallucination Rate, measuring the frequency of mathematically invalid reasoning or unsupported claims; and Faithfulness

Score, evaluating the extent to which solutions adhere to standard mathematical rules, definitions, formulas and - when applicable, retrieved reference materials, with higher scores reflecting more coherent and well-grounded reasoning.

2.5. Experimental results & discussion

This section presents the quantitative experimental results of the proposed mathematics chatbot and discusses the main findings based on the evaluation metrics described in the previous section. The objective is to analyze gains in solution accuracy, mitigation of hallucination in mathematical reasoning, and performance differentials between the proposed system and baseline configurations.

Overall quantitative results

Table 2 summarizes the quantitative evaluation results on the full set of 240 mathematics questions, whose dataset characteristics are detailed in Table 3. The results indicate that the proposed mathematics chatbot consistently outperforms both baseline configurations across all evaluation metrics.

Table 2. Performance comparison between baseline LLMs and the proposed system

Model	Exact Match (EM, %)	Answer Accuracy (%)	Hallucination Rate (%)	Faithfulness Score (0–1)
Baseline LLM (no RAG)	62.4	68.7	21.3	0.61
Baseline LLM + RAG	71.9	77.5	13.8	0.74
Proposed System	**82.6**	**87.9**	**6.4**	**0.89**

The proposed system substantially outperforms both baselines across all evaluation metrics. In terms of Exact Match, it achieves 82.6%, compared to 62.4% for the baseline LLM and 71.9% for LLM + RAG, demonstrating the effectiveness of mathematical constraints and controlled inference in producing accurate final answers for well-defined problems such as arithmetic, algebra, and equation solving. A similar pattern is observed for Answer Accuracy, where the proposed system reaches 87.9%, surpassing 68.7% and 77.5% for the two baselines, indicating improvements not only in final correctness but also in step-by-step reasoning quality.

A key contribution of the proposed approach is its ability to mitigate hallucination in mathematical reasoning. The hallucination rate drops to 6.4%, compared to 21.3% for the baseline LLM and 13.8% for LLM + RAG, highlighting that retrieval alone is insufficient without explicit mathematical constraints. Consistently, the Faithfulness Score increases to 0.89, markedly higher than 0.61 and 0.74, reflecting stronger adherence to standard definitions, formulas, and logical consistency, an essential requirement in educational applications to avoid reinforcing misconceptions.

Further analysis by grade level and question type shows consistent gains across all educational levels, with the largest improvements at Lower Secondary and Upper Secondary levels, where multi-step reasoning is required. For multiple-choice and short-answer questions, gains are mainly reflected in higher Exact Match (EM) scores, whereas for open-ended problems, improvements are more evident in Answer Accuracy and

mitigated hallucination, underscoring the proposed system's robustness in maintaining valid and coherent mathematical reasoning

Table 3. Statistics of the evaluation dataset

	Description
Total number of questions	240
Subject domains	Mathematics, Science, AI Literacy, Ethics
Grade levels	Primary (30%), Lower Secondary (40%), Upper Secondary (30%)
Question types	Multiple-choice (35%), Short-answer (40%), Open-ended (25%)
Average question length	23.4 tokens
Average answer length	86.7 tokens
Evaluation method	Dual human annotation
Inter-annotator agreement	Cohen's $\kappa = 0.82$

2.6. Development direction

Despite the positive results, several directions remain for future work. These include developing more specialized models for complex problem types (e.g., spatial geometry and advanced multi-step reasoning), expanding the mathematical data repository to cover more diverse and richer problem structures, and enhancing visual interpretation of diagrams and graphs. In addition, the system can be extended to support intelligent learning features such as systematic error analysis, adaptive feedback mechanisms, and automatic generation of practice problems.

3. Conclusions

This study presents a comprehensive framework for developing a Vietnamese Mathematics Chatbot based on a large language model integrated with a Retrieval–Augmented Generation architecture. By systematically combining key technical components—including the Qwen3-VL-8B multimodal large language model, the Milvus vector database, the BGE-m3 embedding model, and the LangGraph-based state management mechanism—the proposed system establishes a unified and robust query processing pipeline. This pipeline is capable of handling heterogeneous input modalities and generating accurate, well-structured, and pedagogically aligned mathematical solutions in accordance with the Vietnamese national mathematics curriculum.

Experimental and implementation results demonstrate the effectiveness of the proposed system across multiple dimensions. More precisely, the chatbot exhibits strong capabilities in recognizing mathematical problems from both textual and visual inputs, analyzing problem requirements, selecting appropriate solution strategies, and presenting step-by-step explanations. What is more, the integration of a semantic retrieval mechanism enables the system to access relevant mathematical knowledge efficiently, while the state-based dialogue management framework ensures consistency and coherence across multi-turn interactions. The combination of RAG and large language models effectively leverages the reasoning strengths of LLMs while substantially

mitigating hallucination by grounding responses in a structured knowledge base. These findings confirm that the application of RAG-based architectures in mathematics education is not only feasible but also yields significant practical value.

Beyond the current results, this study identifies several promising directions for future research aimed at further enhancing system performance and applicability. Potential extensions comprise expanding and diversifying the mathematical knowledge base, developing specialized sub-models tailored to distinct mathematical domains, advancing visual interpretation and diagram comprehension capabilities, and incorporating advanced error diagnostics and personalized practice recommendation modules. Such enhancements would enable the system to evolve into a more comprehensive and adaptive mathematics assistant, supporting the development of personalized learning environments and addressing the diverse instructional needs of both students and educators.

REFERENCES

[1] Bengio Y, Courville A & Vincent P, (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828.

[2] Mikolov T, Chen K, Corrado G & Dean J, (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[3] Winkler S & Söllner A, (2022). Unleashing the potential of chatbots in education: A systematic review. *Computers & Education: Artificial Intelligence*, 3, 100074, 2022.

[4] Huang R & et al., (2022). Chatbots for language learning: A meta-analysis. *Educational Research Review*, 37, 100487, 2022.

[5] Druga L & et al., (2017). Growing up with AI: Cognition and creativity in children's interactions with intelligent agents. *Proceedings of the 2017 ACM Interaction Design and Children Conference*, 351-362.

[6] Ji S, Zhang R, Wei B & Ma X, (2023). Survey of hallucination in natural language generation. *arXiv preprint arXiv:2302.0807.*

[7] Nguyen A, Yosinski K & Clune J, (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 427-436.

[8] Lewis P & et al., (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

[9] Karpukhin J & et al., (2020). Dense passage retrieval for open-domain question answering. *EMNLP*, 6769-6781.

[10] Xiao X & et al., (2024). BGE-M3: A multi-function, multi-lingual and multi-granularity text embedding model. *arXiv preprint arXiv:2402.07872*.

[11] Qwen Team, (2023). Qwen-VL: A comprehensive multimodal foundation model. *arXiv preprint arXiv:2308.02949*.

[12] LangChain AI, (2024). LangGraph: State-graph workflows for LLM applications. Available: https://python.langchain.com/docs/langgraph/.