Adaptive Multi-Agent Tutoring AI for Multimodal Mathematics Conversational Learning

Quy Minh Le
ISY Labs
Hanoi, Vietnam
quy.le@isy.com.ai

Hoang D. Nguyen
University College Cork
Cork, Ireland
hn@cs.ucc.ie

Abstract

Providing tailored and multimodal academic support in mathematics remains a complex task, especially when the goal is to deliver timely assistance aligned with specific curricula. Many current question-answering (QA) systems are limited in interactivity, perform inconsistently across different input types (e.g., text, image, voice), and often fail to respond effectively to individual learning needs. These challenges are even more prominent in subjects like algebra or geometry, where abstract reasoning and step-by-step solutions are crucial. This paper introduces a conversational AI tutoring system grounded in Large Language Models (LLMs) and multi-agent design, with each agent addressing a specific instructional function (Informer, Verifier, Insight, Practice, Tutor). The system integrates multimodal inputs and features a personalized feedback loop that tracks student difficulties, updates learning records, and suggests suitable practice tasks or learning videos. Evaluated with ninth-grade students in Vietnam, the system has demonstrated promising results for both educational impact and broader applicability.

CCS Concepts

• **Applied computing → Collaborative learning**; • **Human-centered computing → HCI theory, concepts and models**; • **Computing methodologies → Natural language processing**; **Multi-agent systems**; • **Information systems → Recommender systems**.

Keywords

Multimodal Question Answering, Conversational AI, Intelligent Tutoring Systems, RAG, Multi-Agent Systems, AI in Education, Empirical Study

ACM Reference Format:
Quy Minh Le and Hoang D. Nguyen. 2025. Adaptive Multi-Agent Tutoring AI for Multimodal Mathematics Conversational Learning. In *Proceedings of the 2nd ACM Workshop in AI-powered Question & Answering Systems (AIQAM '25), October 27–28, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3746274.3760396

AIQAM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2056-7/2025/10
https://doi.org/10.1145/3746274.3760396

1 Introduction

The integration of Artificial Intelligence (AI) into education has grown rapidly, especially with the increasing shift toward online and hybrid learning environments. This shift has created a clear need for tools that offer both scalable and personalized academic support. Such tools are intended to complement traditional teaching by enabling faster feedback and personalized learning experiences for students.

Mathematics, however, poses particular challenges for AI-based systems. Its concepts are often abstract, requiring structured, step-by-step reasoning that generic solutions often struggle to provide. Students vary in how they understand mathematical concepts, and generic materials often lack alignment with these diverse learning needs.

General-purpose Large Language Models (LLMs) have shown impressive capabilities, yet their application in education remains limited by issues such as alignment with specific curricula and inconsistent reliability in formal academic contexts. These limitations are even more pronounced in non-English settings. Recent multilingual benchmarks point to persistent issues, including cultural bias, dependence on memorization-based testing formats, and insufficient support for low-resource languages [18]. In practice, these gaps have been shown to result in reduced reasoning performance, as demonstrated in the IRLBench study comparing English and Irish [18].

In response to these limitations, this work presents a multi-agent tutoring system for mathematics learning that supports multimodal inputs (text, speech, and image). Each agent in the system is designed with a dedicated pedagogical role, allowing not just curriculum-aligned answers but also proactive detection and remediation of student weaknesses during the learning process.

The key contributions of this work include:

(1) Introduces a multi-agent tutoring AI framework for mathematics education, in which each LLM agent plays a distinct pedagogical role to enhance clarity and explainability.
(2) Proposes a comprehensive multimodal processing framework for text, speech, and image-based queries, featuring a task-specific multistep RAG approach for image-based queries aimed at improving retrieval accuracy.
(3) Constructs a domain-specific knowledge base aligned with the official Vietnamese ninth-grade mathematics curriculum, combining curated textbook content and educational video metadata.

(4) Presents an empirical study involving real students and a structured evaluation plan, providing early insights into the effectiveness and usability of multimodal AI tutoring systems in real-world educational settings.

The system is also presented as an empirical study to advance the understanding of AI-powered Q&A systems for multimedia [11].

2 Related Work

Our approach builds on advances in three closely related areas: Intelligent Tutoring Systems (ITS), Conversational AI in Education, and Multimodal Question Answering (MMQA).

2.1 Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) have a long and rich history, evolving from early rule-based platforms to more sophisticated, data-driven, and AI-powered systems [1]. A scientometric review by Guo et al. highlights the multidisciplinary nature of ITS research and the recent trend towards integrating modern AI techniques to create more dynamic and natural learning interactions [5]. Our work builds upon this evolution by leveraging a multi-agent structure where each agent is assigned a specific pedagogical role, moving beyond a single monolithic AI model to a more structured and explainable tutoring framework.

2.2 Conversational AI in Education

The application of chatbots in educational settings has been systematically reviewed, with studies covering prominent systems such as Khanmigo, DeepTutor, and NeuroChat [8]. A key finding from this research is that many existing systems are still primarily text-based and often lack the capability to proactively diagnose student knowledge gaps. Another comprehensive review by Labadze et al. notes that although widely used AI chatbots — such as GPT, Ada, and Socratic — have enabled more interactive natural language environments, they typically lack alignment with specific curricula and do not consistently offer adaptive support based on individual student performance [9]. Our system directly addresses these gaps through its proactive `Insight` and `Practice` agents, which are designed to detect and remediate specific weaknesses based on the ongoing conversation.

2.3 Multimodal Question Answering (MMQA)

Recently, the field of MMQA has seen a significant surge in multi-agent approaches for handling complex and diverse queries. Research has demonstrated the power of specialized agents in synthesizing information from diverse sources, including document understanding [6], video QA [7], and attribution-based question answering leveraging Retrieval-Augmented Generation (RAG) methods [2]. Rajput et al. also propose a multi-agent perspective that improves information synthesis in MMQA [13].

A recent and relevant study is IRLBench [18], a multimodal benchmark developed from Irish national examination materials. Although both projects emphasize the need for culturally and pedagogically grounded evaluation resources for low-resource languages, their goals differ significantly. IRLBench focuses on benchmarking open-ended generative reasoning from complex PDF documents, whereas our system is designed as a fully interactive tutoring platform. In particular, our approach includes a dynamic and personalized pedagogical feedback loop — implemented via the `Insight` and `Practice` agents — which extends beyond the static evaluation purpose of conventional benchmarks.

However, most of these advanced works focus primarily on general-purpose datasets and benchmarks. To the best of our knowledge, no empirical study has yet concentrated on applying a multi-agent, multimodal QA system to the specific domain of mathematics tutoring grounded in a national curriculum. Our work aims to fill this void by presenting a complete empirical study that combines a multistep RAG process with a framework of agents assigned with clear pedagogical roles, tailored specifically to the Vietnamese ninth-grade mathematics curriculum.

3 Methodology

Our system is designed based on a multi-agent architecture [17] to handle the complex and dynamic nature of a tutoring conversation. It is implemented as a web application using the Streamlit framework [14], providing a lightweight and interactive interface. The system is built on Haystack, a widely used open-source toolkit that supports the development of applications using large language models (LLM) [3], due to its modular design and built-in support for retrieval-augmented generation pipelines. The overall architecture, depicted in Figure 1, is composed of a central **Orchestration Hub** that acts as the main interface to the user, and two specialized functional modules: a **Problem-Solving Engine** for providing accurate answers, and a **Learning Engine** for driving personalization.

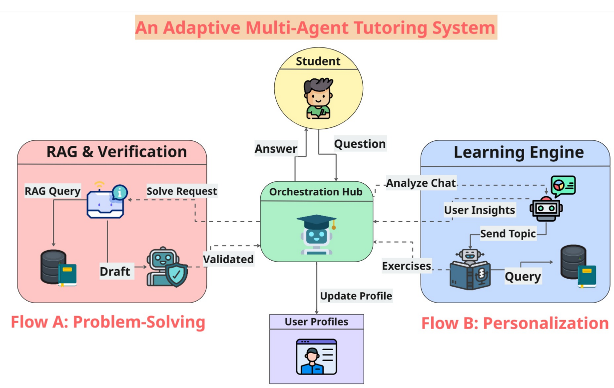

Figure 1: The proposed adaptive multi-agent architecture. The central Orchestration Hub receives user queries and delegates tasks to two specialized modules: the RAG & Verification engine for problem-solving (Flow A), and the Learning Engine for personalization (Flow B).

3.1 User Interaction and Interface

The system provides a dynamic, multimodal conversational interface that enables users to engage with the underlying multi-agent architecture seamlessly.

As shown in Figure 2, the interface supports various input modes: students can type a question, upload an image, or submit a spoken query via speech-to-text. This design accommodates diverse user preferences and problem formats, particularly those with visual components, which regularly appear in school-level mathematics problems.

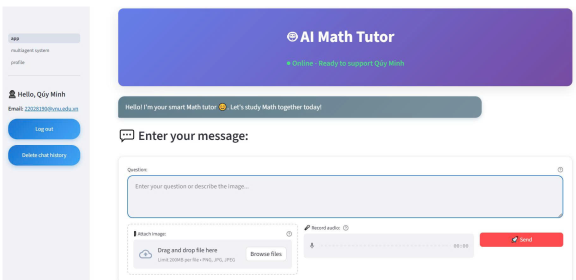

Figure 2: The main conversational interface with multimodal input options (text, image, and speech).

A key feature of our system is its ability to provide proactive and personalized feedback. As shown in Figure 3, after a series of interactions in which the system detects a potential knowledge gap, the Learning Engine is activated. The chatbot then automatically initiates a new conversational turn, suggesting tailored practice exercises and learning resources to address the specific misconception identified.

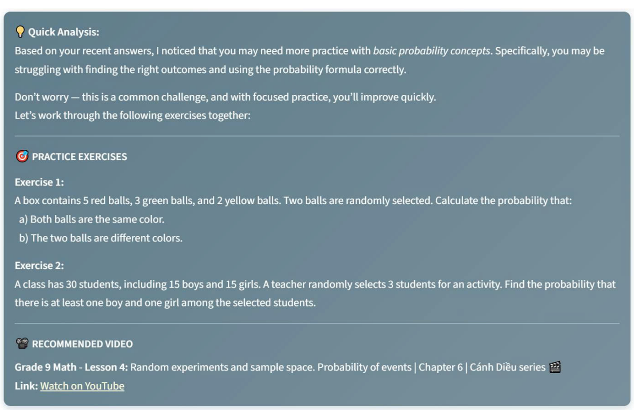

Figure 3: The system proactively identifies a misconception and suggests personalized exercises.

The insights generated from these interactions are persistently stored and visualized in the student's personal learning profile. This profile page serves as a dashboard for students to track their progress and understand their knowledge gaps.

Figure 4 presents a summary of the student's progress, including their most recently identified misconception and the total number of concepts identified for further improvement. This overview offers learners a clear and immediate understanding of their current academic progress.

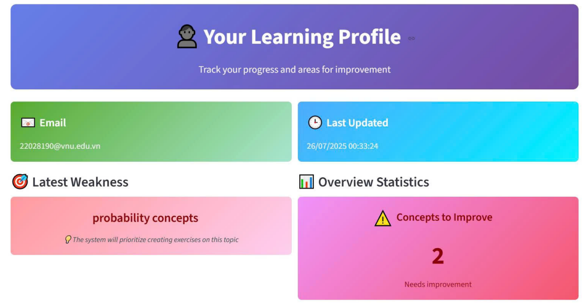

Figure 4: The summary view of the learning profile, highlighting the most recent weakness and overall statistics. This card-based layout provides a quick overview for the student.

Furthermore, the profile offers a detailed breakdown of all identified weaknesses. Figure 5 displays the cumulative list of misunderstood concepts, which is appended over time, creating a long-term learning history. This list is supplemented with actionable learning suggestions that help students address specific knowledge gaps. This entire process creates a continuous and transparent feedback loop, empowering students to take control of their learning journey.

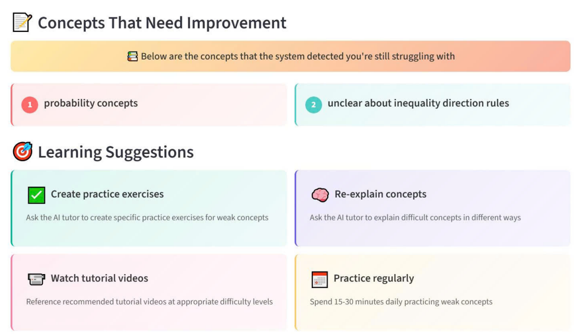

Figure 5: The detailed view of the learning profile, showing the cumulative list of all misunderstood concepts and providing actionable learning suggestions.

3.2 The Multi-Agent Framework

The core reasoning capabilities of the system are distributed among five specialized agents, each designed to fulfill a specific instructional role. This modular architecture increases the fault tolerance of the system and improves the interpretability of its responses.

3.2.1 Intent Classification for Dynamic Orchestration. To support dynamic, context-sensitive dialogue, the tutoring module relies on an intent classification mechanism. This enables the system to identify user goals and route queries to the relevant agent or initiate a targeted conversational pathway. Intent recognition is handled by the language model using a minimal-example prompting strategy, in which the model is guided by representative samples for each predefined intent to ensure reliable classification.

We designed six key intents, categorized into two main groups:

Core Academic Intents. This group encompasses essential learning interactions commonly encountered in mathematics tutoring scenarios.

- **math_question:** Activated when students ask about mathematical concepts, request definitions, or seek step-by-step solutions. *Example: "solve the equation x^2 - 5x + 6 = 0."*
- **request_for_practice:** Recognizes situations where students explicitly request additional exercises to strengthen their understanding. *Example: "give me more exercises on this topic."*

Communication and Support Intents. This group manages the conversational and affective aspects of the tutoring experience.

- **study_support:** Addresses general questions about learning strategies and motivation. *Example: "how can I study math better?".*
- **expression_of_stress:** Detects expressions of frustration or fatigue, allowing the system to switch to an empathetic, supportive role. *Example: "this problem is too hard, I'm really discouraged".*
- **greeting_social:** Manages common social interactions like greetings and thanks to make the conversation feel more natural. *Example: "hi there!" or "thanks for your help!"*
- **off_topic:** Catches queries unrelated to mathematics or learning, enabling the chatbot to politely redirect the conversation. *Example: "do you know who won the football match last night?"*

3.2.2 Agent Roles and Responsibilities.

To effectively emulate a human tutoring experience, the system's tasks are deconstructed and assigned to agents with specific pedagogical roles. These agents are organized into two primary functional engines, which are coordinated by the central `Tutor Agent`. This separation of concerns allows the system to handle different aspects of the conversation — such as factual instruction, personalized feedback, and social interaction — with specialized logic.

Tutor Agent (Orchestration Hub). This agent serves as the central brain of the system and the sole point of contact with the user. It is responsible for managing the overall conversational flow, invoking the intent classifier, and delegating tasks. As shown in Figure 1, when a user asks a 'Question', the Hub sends a 'Solve Request' to the Problem-Solving Engine. After a series of interactions, it proactively triggers the Learning Engine by sending the conversation history for analysis ('Analyze Chat'). Then it receives processed outputs from both engines ('Validated' answers or 'Exercises') and delivers them to the student. It is also responsible for the 'Update Profile' action, persisting insights to the database.

Problem-Solving Engine (Informer & Verifier Agents). This engine (Flow A in Figure 1) is designed to generate accurate, verified, and curriculum-aligned responses to mathematical queries. It operates as an independent module triggered by the `Tutor Agent` in response to any `math_question` intent, which sends a 'Solve Request' as input.

- The **Informer Agent** initiates the RAG process. It receives the student's query and conversation history, performs a 'RAG Query' on the curated knowledge base, and constructs a detailed, step-by-step 'Draft' response.
- This draft is forwarded to the **Verifier Agent**, which serves as a mathematical validation layer. The draft is assessed for mathematical accuracy, logical consistency, and alignment with the original query. If necessary, it suggests corrections based on identified issues.

The resulting output — a 'Validated' and, if required, revised response — is then returned to the `Tutor Agent` for presentation to the student.

Learning Engine (Insight & Practice Agents). This engine (Flow B in Figure 1) supports personalized adaptation by analyzing interaction patterns to discover potential learning challenges. Activation occurs when the `Tutor Agent` observes recurring signals that indicate conceptual confusion, proactively triggering the engine with an 'Analyze Chat' signal containing the conversation history.

- The **Insight Agent** functions as a diagnostic component. It silently analyzes the latest sequence of exchanges (conversation history) to identify potential areas of misunderstanding or student weaknesses. This is achieved by prompting the LLM with the conversation history and a set of illustrative examples (a few-shot technique) and instructing it to identify specific mathematical concepts that the student appears to be struggling with. When a knowledge gap is detected, its findings — referred to as 'User Insights' — are formatted into a structured JSON object, which includes the field `misunderstood_concepts` listing topics that require further attention.
- The JSON output follows two pathways. One copy is persistently stored in the learner's profile by the `Tutor Agent` for longitudinal tracking (see Section 3.2.3). At the same time, the **'Insight Agent'** uses the same information to 'Send Topic' to the **Practice Agent** for targeted remediation.
- The **Practice Agent** addresses the gap by taking two specific actions: (1) generating original practice 'Exercises' designed to reinforce the relevant concept, and (2) retrieving an instructional video recommendation by performing a 'Query' on indexed metadata.

Once completed, the feedback package — comprising both exercises and video materials — is routed back through the `Tutor Agent`, who provides the content as a learning intervention tailored to the learner and delivered at the right time.

3.2.3 User Profile Persistence.

To enable long-term personalization, the system uses the Supabase framework [15] to maintain a persistent profile for each user. The `user_profiles` database stores data generated by the `Insight Agent`, such as misunderstood concepts. This allows the system to have a "long-term memory" adapting its suggestions and support across multiple learning sessions. Row-level security policies are enforced to guarantee the privacy of each student's learning data.

3.3 Multimodal Mathematical Conversational AI

3.3.1 Dataset Curation.

The system's knowledge base is grounded in the ninth-grade mathematics curriculum of Vietnam. This was selected as a critical milestone for the progression of Vietnamese K-12 students.

- **Textbook Data:** The full content of the two-volume textbook set was extracted using the MinerU framework [19]. The raw text was then cleaned using a sequence of processing steps to eliminate elements such as page numbers and headers. This produced a clean `.txt` file that serves as input to the RAG system.
- **Video Metadata:** We created a "Video Cheatsheet" by curating metadata from a popular educational YouTube playlist by Vietjack covering the same curriculum. This `videos.json` file contains titles, keywords, and summaries for each video, enabling targeted recommendations.

3.3.2 Multimodal Input Processing. The system is designed to handle three types of user input, with all generative tasks powered by the Gemini 1.5 Pro model [4]:

- **Text**: Textual queries are processed through a standard RAG flow, where the query is used to retrieve relevant text chunks that augment the LLM's context.
- **Speech**: Voice inputs are first transcribed into text using the `faster-whisper` library [16]. This initial step converts the audio input into normalized text, enabling compatibility with the existing text-based processing pipeline.
- **Image**: To handle image-based queries, we implement a structured workflow RAG process that begins with a crucial pre-retrieval query transformation step. This approach is motivated by the challenges of processing complex documents, such as exam papers or textbooks. Following the approach adopted in recent benchmark construction efforts — notably IRLBench [18] — we employ the Gemini-Vision model as an advanced OCR tool in the initial stage to extract textual context from images. This extracted text is then combined with the user's prompt to form an enriched query, which is subsequently used in the RAG retrieval stage to find relevant documents.

3.3.3 Utilization of Video Metadata for Learning Recommendations. Our system leverages multimedia data not only to handle user queries, but also as an integral feedback resource, directly supporting one of the central objectives of the AIQAM workshop. The `Practice Agent`, upon receiving a weakness topic from the `Insight Agent`, queries the `videos.json` dataset to find the video whose title, keywords, or summary best matches the identified topic. This enables the chatbot to give precise recommendations, such as: "I notice you might need more practice with Vieta's formulas. You can refer to the video 'Grade 9 Math - Lesson 3: Vieta's formulas' in the playlist." This approach demonstrates a practical and effective method for achieving the stated objective of "utilization of multimedia data for providing valuable answers and feedback."

4 Evaluation

4.1 User Group

To evaluate the effectiveness of the system, we conducted an initial empirical study with 10 ninth-grade students, aged 15-16, who were randomly selected from various secondary schools in Hanoi, Vietnam. Although this sample size is intended for preliminary insight and not statistical generalization, the participants represent a typical demographic for the system's target audience.

4.2 Evaluation Procedure

The evaluation will be conducted using a Google form designed to collect user feedback. We will employ a task-based evaluation method in which students are asked to complete specific tasks such as solving a problem from an image, triggering the proactive learning loop, and checking their user profile.

4.3 Evaluation Criteria

The evaluation form includes both quantitative metrics (on a 5-point Likert scale) and open-ended questions for qualitative feedback. We assess five main dimensions: Usefulness, Accuracy, Personalization, User Experience, and Safety.

4.4 Preliminary Observations

During internal testing, we observed that the multi-step RAG architecture for image-based queries produced significantly more accurate and relevant results than relying solely on the user's textual description. This finding highlights the importance of transforming the query before the retrieval phase, which is found to play a critical role in enabling effective multimodal QA.

5 Results, Discussions and Ethical Considerations

5.1 Results

The evaluation involved 10 ninth-grade students from various lower secondary schools in Hanoi. A significant majority (80%) had prior experience with general-purpose AI chatbots, citing a lack of accuracy and the inability to solve curriculum-specific problems as their primary weaknesses with existing tools. This confirms the motivation for our specialized system.

The quantitative results of the user evaluation are summarized in Table 1, which groups the metrics by the core capabilities of the system.

The performance of the system **core QA performance** was rated highly. Text-based interactions, including the accuracy of solutions and contextual understanding in follow-up questions, achieved strong average scores of 4.7 and 4.6 out of 5, respectively. The image understanding capability, a key feature of our multi-step RAG architecture, was also rated very favorably at 4.4. As anticipated, the speech-to-text functionality received a lower score of 3.8 due to real-world technical and accuracy challenges, and is identified as a primary area for future technical refinement.

The most significant results were observed in the **Proactive Personalization Loop**. This set of features was the highest-rated aspect of the entire system. The system's ability to automatically identify student weaknesses achieved an exceptional score of 4.7. Crucially, the relevance of the dynamically generated exercises and the usefulness of the curriculum-aligned video recommendations were also rated very highly, with scores of 4.7 and 4.8, respectively.

Finally, the system's ability to provide a safe and supportive environment was validated. The composite **User Experience & Safety** metric, which includes politeness, empathy, and safety, achieved a strong score of 4.4, which confirms the effectiveness of the `Tutor Agent`'s conversational design.

Table 1: Average User Ratings for System Capabilities (Scale 1-5, N=10)

Capability	Average Score
Core QA Performance	
Accuracy of Math Solutions (Text)	4.7 / 5
Contextual Understanding (Text)	4.6 / 5
Image Understanding	4.4 / 5
Speech Understanding	3.8 / 5
Proactive Personalization Loop	
Proactive Weakness Detection	4.7 / 5
Relevance of Generated Exercises	4.7 / 5
Usefulness of Video Recommendations	4.8 / 5
User Experience & Safety	
Affective Support & Safety	4.4 / 5

5.2 Discussion

The results of our empirical study strongly validate our multi-agent, multimodal approach for personalized tutoring. The high scores on accuracy and contextual understanding (Table 1) directly address the primary weaknesses identified by participants in general-purpose chatbots. By grounding the system's knowledge in a curated, curriculum-specific dataset via RAG, we significantly increased user trust, a critical factor in educational applications.

One notable observation from the evaluation is the positive reception of the proactive personalization loop. Feedback from participants indicated that the `Insight` and `Practice` agents contributed meaningfully to the learning experience by identifying student weaknesses and offering tailored practice. This aligns with the intended goal of moving beyond static Q&A and toward more adaptive, personalized tutoring. In particular, the use of video metadata for curriculum-aligned recommendations was frequently mentioned in qualitative feedback as a distinctive feature that increased the perceived level of personalization.

Furthermore, the favorable evaluation of the image-based problem-solving task serves as a strong endorsement for our multi-step RAG architecture. This approach, which uses a vision model for a pre-retrieval query transformation step, is similar in principle to the data extraction pipelines used in recent benchmark creations like IRLBench [18]. Our results demonstrate that this method is effective not only for data set creation, but also for real-time query understanding in an interactive system, successfully handling requests with minimal textual context.

Although this work focuses on the Vietnamese educational context, it reveals challenges that reflect broader trends in multilingual AI research. The IRLBench study [18], for example, illustrates a notable disparity in reasoning capabilities between English — a high-resource language — and Irish, a low-resource one. This observation echoes the intrinsic challenges of developing reliable AI systems for underrepresented languages such as Vietnamese. Through this case analysis, we add to the growing discourse on the real-world obstacles and design requirements necessary for culturally aware, educational AI solutions in low-resource linguistic contexts.

5.3 Limitations

The primary limitation of our empirical study is the small sample size **(N = 10)**. Although feedback provides valuable initial information, the results are not statistically significant and cannot be generalized. A larger-scale study, potentially including a control group or comparison with human tutors, is necessary to draw more robust conclusions about the system's educational impact.

Our system's performance is also heavily dependent on the quality of a single underlying LLM (Gemini 1.5 Pro). As demonstrated by the IRLBench study that benchmarked multiple models [18], performance can vary substantially between different foundational models. Additionally, the `Insight Agent`'s ability to correctly identify misconceptions is a complex reasoning task and may not always be accurate.

The current retrieval-augmented generation setup, while effective for this study, also presents opportunities for further refinement. The document analysis phase relies on a fixed-size segmentation strategy, which may not always preserve the full semantic context of complex theorems. The retrieval process uses only dense vector search, which can sometimes struggle with queries containing specific keywords or technical terms. Finally, the system lacks a re-ranking mechanism to further refine and diversify the retrieved documents before generation.

5.4 Ethical Considerations

We designed the system with several ethical considerations in mind to ensure a safe and responsible learning environment.

- **Emotional Handling:** The `Tutor Agent` is explicitly prompted to respond safely and empathetically when it detects an `expression_of_stress` intent. It is instructed to validate the student's feelings and suggest simple, safe actions like taking a break, while clearly refraining from offering any professional psychological advice to ensure student well-being.
- **Knowledge Boundaries:** The chatbot is designed to politely refuse `off_topic` questions. This mechanism ensures that the conversation remains focused on the educational context and prevents exposure to potentially inappropriate content for a young audience.
- **Data Privacy:** We use Supabase with row-level security policies to ensure that each student's learning profile, which contains information about their academic weaknesses, is kept strictly private and accessible only to them. All data collected are anonymized and used solely for research purposes.

6 Conclusion and Future Work

In this paper, we presented an adaptive multi-agent system for multimodal mathematics tutoring. Our empirical study demonstrates a viable architecture that combines a specialized agent framework with a multi-step RAG process to provide a personalized and curriculum-aligned learning experience. The positive results from our user evaluation confirm the value of a proactive, personalized feedback loop in an educational AI system.

Future work will focus on several key areas. One priority is to improve the visual processing capabilities of the system, including handwriting recognition from student worksheets. Currently, we will address the limitations of the speech-to-text feature identified in our evaluation by exploring more robust transcription models or fine-tuning techniques to improve its accuracy and reliability. We also aim to expand the knowledge base to other subjects such as Physics and Chemistry.

Furthermore, we envision evolving our RAG pipeline into a more advanced reasoning engine. In the initial phase of input handling, we plan to apply refined segmentation strategies, such as semantic segmentation, to produce logically consistent knowledge units [20]. To strengthen retrieval robustness, the system will integrate a hybrid search method combining dense and sparse representations, followed by an intelligent reranking procedure that considers both relevance and diversity [10, 21]. Finally, to support extended dialogues and more intricate problem scenarios, we intend to examine prompt compression approaches for managing the language model's context window more efficiently — an area addressed in several recent studies on multimodal RAG [12].

Finally, integrating a gamification system informed by user profile data could further increase student engagement. We will also continue to refine the user evaluation process to gather more comprehensive data. Moreover, future evaluations will aim to incorporate more objective performance metrics, such as pre-test/post-test scores and response times, and explore automatic evaluation pipelines to complement user feedback. We also intend to benchmark the system's performance using different foundational LLMs to assess the generalizability of our architectural findings.

Acknowledgments

This publication has emanated from research supported in part by a grant from Research Ireland under Grant number [12-RC-2289-P2] which is co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

We would like to thank the students who will participate in our user study for their valuable feedback.

References

[1] Ali Alkhatlan and Jugal Kalita. 2018. Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments. arXiv:1812.09628 [cs.HC] https://arxiv.org/abs/1812.09628

[2] Ines Besrour, Jingbo He, Tobias Schreieder, and Michael Färber. 2025. RAGentA: Multi-Agent Retrieval-Augmented Generation for Attributed Question Answering. arXiv:2506.16988 [cs.IR] https://arxiv.org/abs/2506.16988

[3] deepset GmbH. 2023. Haystack: End-to-end framework for question answering and retrieval-augmented generation. https://github.com/deepset-ai/haystack. Accessed: 25 July 2025.

[4] Google DeepMind. 2023. Gemini: Google's Multimodal AI Models. https://deepmind.google/models/gemini/. Accessed: 25 July 2025.

[5] Lu Guo, Dong Wang, Fei Gu, Yazheng Li, Yezhu Wang, and Rongting Zhou. 2021. Evolution and trends in intelligent tutoring systems research: a multidisciplinary and scientometric view. *Asia Pacific Education Review* 22, 3 (September 2021), 441–461. doi:10.1007/s12564-021-09697-7

[6] Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. MDocAgent: A Multi-Modal Multi-Agent Framework for Document Understanding. arXiv:2503.13964 [cs.LG] https://arxiv.org/abs/2503.13964

[7] Noriyuki Kugo, Xiang Li, Zixin Li, Ashish Gupta, Arpandeep Khatua, Nidhish Jain, Chaitanya Patel, Yuta Kyuragi, Yasunori Ishii, Masamoto Tanabiki, Kazuki Kozuka, and Ehsan Adeli. 2025. VideoMultiAgents: A Multi-Agent Framework for Video Question Answering. arXiv:2504.20091 [cs.CV] https://arxiv.org/abs/2504.20091

[8] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies* 28, 1 (2023), 973–1018. doi:10.1007/s10639-022-11177-3

[9] Lasha Labadze, Maya Grigolia, and Lela Machaidze. 2023. Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 56. doi:10.1186/s41239-023-00426-1

[10] Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N. Ioannidis, Huzefa Rangwala, and Christos Faloutsos. 2025. HybGRAG: Hybrid Retrieval-Augmented Generation on Textual and Relational Knowledge Bases. arXiv:2412.16311 [cs.LG] https://arxiv.org/abs/2412.16311

[11] Tan Tai Mai, Allie Tran, Quang-Linh Tran, An Nguyen, Hoang D. Nguyen, Tho Quan, Duc-Tien Dang-Nguyen, and Cathal Gurrin. 2025. AIQAM'25: The 2nd ACM Workshop on AI-powered Question Answering Systems for Multimedia. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)* (Dublin, Ireland). Association for Computing Machinery.

[12] Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2025. A Survey of Multimodal Retrieval-Augmented Generation. arXiv:2504.08748 [cs.IR] https://arxiv.org/abs/2504.08748

[13] Krishna Singh Rajput, Tejas Anvekar, Chitta Baral, and Vivek Gupta. 2025. Rethinking Information Synthesis in Multimodal Question Answering A Multi-Agent Perspective. arXiv:2505.20816 [cs.CL] https://arxiv.org/abs/2505.20816

[14] Streamlit Inc. 2022. Streamlit: The fastest way to build and share data apps. https://streamlit.io. Accessed: 25 July 2025.

[15] Supabase Inc. 2022. Supabase: The open source Firebase alternative. https://supabase.com. Accessed: 25 July 2025.

[16] SYSTRAN. 2023. faster-whisper: A reimplementation of OpenAI's Whisper model using CTranslate2. https://github.com/SYSTRAN/faster-whisper. Accessed: 25 July 2025.

[17] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322* (2025).

[18] Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2025. IRLBench: A Multi-modal, Culturally Grounded, Parallel Irish-English Benchmark for Open-Ended LLM Reasoning Evaluation. arXiv:2505.13498 [cs.CL] https://arxiv.org/abs/2505.13498

[19] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. MinerU: An Open-Source Solution for Precise Document Content Extraction. arXiv:2409.18839 [cs.CV] https://arxiv.org/abs/2409.18839

[20] Jihao Zhao, Zhiyuan Ji, Yuchen Feng, Pengnian Qi, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025. Meta-Chunking: Learning Text Segmentation and Semantic Completion via Logical Perception. arXiv:2410.12788 [cs.CL] https://arxiv.org/abs/2410.12788

[21] Tong Zhou. 2025. Knowledge-Aware Diverse Reranking for Cross-Source Question Answering. arXiv:2506.20476 [cs.CL] https://arxiv.org/abs/2506.20476