Speech-RAG: A Multimodal Course Assistant with Phi-4-MM and Qwen3-Embedding

Chaoren Wang (122090513)
chaorenwang@link.cuhk.edu.cn

Abstract

In the era of large language models (LLMs), students still face challenges in accessing course-specific information efficiently, especially in hands-free scenarios. General-purpose models often hallucinate critical details like classroom locations or exam weights. This report presents "Speech-RAG": a course assistant designed for persistent context retention. By integrating Microsoft's Phi-4-MM for multimodal speech understanding and Qwen3-Embedding for specialized retrieval-augmented generation (RAG), our system eliminates the fragmentation between document analysis and voice interaction. Validated against a synthetic audio test set generated by Google Gemini TTS, our system achieves 79% accuracy on spoken queries compared to 58% for the baseline, effectively reducing factual errors. This improvement highlights the critical role of retrieval augmentation in mitigating hallucinations for course-specific information. The source code will be publicly available at `https://github.com/yuantuo666/CSC3160-speech_rag`.

1 Introduction

In university settings, students frequently require immediate access to information in hands-free scenarios, asking questions such as "Where is today's CSC3160 lecture held?" Our benchmarking reveals that general-purpose large models (No-RAG) often provide confident but incorrect answers—a phenomenon known as hallucination [1]. For instance, a model might incorrectly state "SDS Office" instead of the correct "TB102" classroom, or misquote the final exam weight as 30% when it is actually 45%. Such "confident errors" can lead to serious real-world consequences, such as missing a class or misallocating study efforts.

Furthermore, existing tools are often fragmented. While Retrieval-Augmented Generation (RAG) addresses accuracy, mainstream platforms frequently separate "document analysis" from "voice conversation," failing to meet the need for immediate, hands-free queries in mobile scenarios.

We propose a course voice assistant with memory capabilities. It features a pre-loaded Persistent Knowledge Base, eliminating the inefficiency of repetitive file uploads, and achieves a seamless fusion of voice interaction with high-precision RAG.

Key Contributions This work makes several contributions to the field of multimodal educational technology. First, our system achieves significant **precision correction**, rectifying over 20% of factual errors made by general-purpose models on queries related to course information, such as attendance policies and assignment deadlines. Second, we demonstrate a novel **full-stack integration** of Microsoft's Phi-4-MM[1] for speech understanding with vLLM [2] for accelerated inference, creating a seamless pipeline that bridges the gap between document analysis and voice interaction. Third, by leveraging **specialized retrieval** with Qwen3-Embedding[2] [3], the system gains a deep understanding of domain-specific terminology, enabling it to accurately answer complex questions like "Which lectures does Assignment 3 cover?" by locating precise information rather than guessing.

[1] `https://huggingface.co/microsoft/Phi-4-multimodal-instruct`
[2] `https://huggingface.co/Qwen/Qwen3-Embedding-0.6B`

Finally, we establish the system's robustness through **real-world validation**, using a synthetic dataset generated by Google Gemini TTS (`gemini-2.5-flash-preview-tts`) [4] to simulate diverse speaker accents and emotional tones, ensuring reliable performance in practical scenarios.

2 Related Work

Retrieval-Augmented Generation (RAG): RAG enhances LLMs by retrieving relevant documents to ground responses in factual evidence [5]. While early works focused on open-domain QA, recent advancements have explored integrating RAG with speech processing to create end-to-end spoken question-answering systems. Our work builds on this by applying RAG specifically to the domain of course management with multimodal inputs.

Multimodal LLMs: The landscape of Multimodal LLMs is rapidly evolving. Models like Phi-4-MM [6] and Qwen3-Omni [7] have demonstrated impressive capabilities in processing audio natively alongside text. Our system leverages Phi-4-MM's ability to handle speech directly, improving inference speed.

Embedding & TTS: Effective retrieval relies on high-quality text embeddings. The Massive Text Embedding Benchmark (MTEB) [8] highlights the performance of models like Qwen3-Embedding in semantic search tasks. On the synthesis side, Neural TTS advances (e.g., F5-TTS [9], CosyVoice2 [10]) allow for the generation of highly naturalistic speech, which we utilize via Gemini TTS [4] to create a challenging and realistic evaluation dataset.

3 Approach

Our system implements a comprehensive pipeline that transforms audio input into accurate, context-aware text responses. The architecture is designed for fast inference and high accuracy.

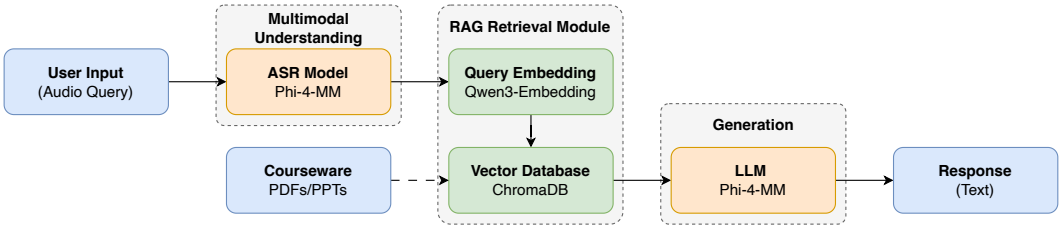

Figure 1: System Architecture of the Speech-RAG Course Assistant. The pipeline integrates multimodal speech understanding with a specialized retrieval module.

3.1 System Pipeline

To address the challenge of providing accurate, hands-free access to course information, we designed a pipeline that integrates multimodal speech understanding with specialized retrieval-augmented generation. The system architecture transforms raw audio input into precise, context-grounded responses through three coordinated stages: multimodal processing, semantic retrieval, and context-aware generation.

The process begins with the **Multimodal Understanding** stage. Upon receiving a voice query, the system utilizes **Phi-4-MM** to process the audio signal. Phi-4-MM generates direct speech interpretation and high-fidelity transcription, preserving the semantic nuance of the student's question.

Following transcription, the **Retrieval Module** queries in the specific course content. The query is encoded into a dense vector representation using **Qwen3-Embedding-0.6B** within the **LangChain**[3] framework. This embedding is used to search **ChromaDB**[4], a vector database containing indexed

[3] `https://github.com/langchain-ai/langchain`
[4] `https://github.com/chroma-core/chroma`

chunks of course materials (PDFs and PPTs). This step ensures that the system identifies the specific lecture segments relevant to the user's inquiry.

Finally, the **Generation** stage synthesizes the answer. We construct a composite prompt that combines the retrieved course context, and the user's original query. This prompt is processed by the **vLLM** inference engine to generate a concise, factually accurate text response.

3.2 Implementation & Optimization

The system is built on a Python stack, leveraging LangChain for workflow orchestration, vLLM for high-throughput inference, and ChromaDB for efficient vector storage. To maximize performance and retrieval accuracy, we focused on two key optimizations. First, for the **embedding model**, we use Qwen3-Embedding-0.6B, which have great semantic capture of specialized course terminology with limited size. Second, to ensure the generation speed, we deploying the model on a remote server accelerated with vLLM and accessed via a Cloudflare Tunnel. This architecture offloads heavy computational tasks from the local device, ensuring a responsive user experience.

4 Experiments

4.1 Experimental Setup

Our experimental design was structured to ensure a robust and realistic evaluation of the system's performance. The foundation of our experiment was the **Knowledge Base**, which was constructed by integrating all course-related PDF handouts and PowerPoint slides into a unified vector store. This repository served as the single source of truth for all retrieval operations.

To create a challenging **Test Set**, we employed **Gemini Flash** (`gemini-flash-latest`) to automatically generate a series of multiple-choice questions (MCQs) and their corresponding correct answers based on the course materials. These text-based queries were subsequently synthesized into high-fidelity audio files using **Gemini TTS** (`gemini-2.5-flash-preview-tts`) with random speakers to ensure speaker diversity. This process yielded a realistic audio test set that captures variations in student intonation, allowing for a rigorous assessment of the system's performance.

4.2 Evaluation Metrics

To quantify the system's effectiveness, we established **Response Accuracy** as the primary evaluation metric. The accuracy was determined by comparing the system's generated answers against the ground truth (GT) derived from the source materials. This comparison was performed automatically using a large language model xFinder[11] to extract answer and compare it with GT. To isolate the impact of different components, we benchmarked performance across four distinct scenarios: (1) **No-RAG + Text**: a baseline using pure text queries without RAG, (2) **No-RAG + Audio**: spoken audio queries without RAG, (3) **Speech-RAG + Text**: pure text queries enhanced with RAG for ablation, and (4) **Speech-RAG + Audio**: our proposed method, which combines audio input with RAG. This comparative analysis allowed us to measure the specific contributions of both the speech recognition and retrieval-augmented generation modules.

Table 1: Performance Comparison across different modalities and retrieval settings (N=100).

Method	Modality	Accuracy (%)	Total Time (s)
No-RAG	Text	66.00	117.19
No-RAG	Audio	58.00	288.16
Speech-RAG (Ablation)	Text	91.00	92.01
Speech-RAG (Ours)	**Audio**	**79.00**	677.58

The test results on 100 queries demonstrate that RAG is the critical factor for accuracy in course queries. As shown in Table 1, incorporating retrieval mechanisms improved performance by 25 percentage points for text inputs (66% vs. 91%) and 21 percentage points for audio inputs (58% vs. 79%). While the transition from text to speech input introduces a performance drop due to the

complexities of acoustic processing, our end-to-end Speech-RAG system (79%) still significantly outperforms the baseline text-only model without RAG (66%). This confirms that even with the noise inherent in speech understanding, grounding the model in specific course materials enables it to correct the majority of hallucinations.

4.3 Case Studies

To illustrate the practical impact of our RAG-based approach, we analyzed two representative cases where general-purpose models are prone to failure. The outcomes are presented below.

Case 1: Lecture Location

User Query:
"What is the location for the Wednesday and Friday lectures?"

No-RAG Response:
"C: SDS office."

Speech-RAG Response:
"B: TB 102."

Case 2: Grading Scheme

User Query:
"What is the max percentage weight of the Final Exam?"

No-RAG Response:
"A: 30 percent."

Speech-RAG Response:
"C: 45 percent."

These examples highlight how the RAG mechanism prevents the model from providing plausible but false information by grounding its response in course documents. The system retrieves correct details, demonstrating its ability to handle factual and numerical data, which are common failure points for non-retrieval models.

5 Conclusion

This project successfully demonstrates a Speech-RAG Course Assistant that addresses the limitations of general AI models. By combining the multimodal capabilities of Phi-4-MM with the precise retrieval of Qwen3-Embedding, we achieved a robust system capable of understanding spoken queries and providing factually accurate responses based on course materials. The system achieves 79% accuracy on spoken queries, significantly outperforming the 58% baseline of general-purpose models. Future work could focus on further improving the first package's speed for real-time conversational speeds and expanding the knowledge base to support multi-course queries simultaneously.

References

[1] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[2] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.

[3] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and

Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[6] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

[7] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[8] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia, 2023. Association for Computational Linguistics.

[9] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

[10] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Huadai Liu, Zhengyan Sheng, Yue Gu, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[11] Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu li, Feiyu Xiong, Bo Tang, and Ding Chen. xfinder: Large language models as automated evaluators for reliable evaluation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Representation Learning*, volume 2025, pages 59850–59892, 2025.