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

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

1 Introduction

University students frequently require hands-free access to course logistics. However, our benchmarking shows that general-purpose models often suffer from hallucination [1], providing confident but incorrect answers regarding venues or grading that can lead to missed classes. Furthermore, existing tools often fragment document analysis and voice interaction, failing to support immediate mobile queries. We propose a course voice assistant that integrates a pre-loaded Persistent Knowledge Base with high-precision RAG, offering a seamless, hands-free solution. The source code is available at `https://github.com/yuantuo666/CSC3160-speech_rag`.

Key Contributions We advance multimodal course assistant through three contributions. First, our system achieves precision correction, reducing factual errors in course queries by over 20%. Second, we implement a integration of Phi-4-MM[1] (with vLLM [2] inference engine) and specialized retrieval via Qwen3-Embedding[2] [3] to accurately locate domain-specific answers. Finally, we validate using a synthetic dataset from Gemini TTS [4], confirming robustness across diverse accents and tones.

2 Related Work

Multimodal LLMs The landscape of Multimodal LLMs is evolving rapidly, with models like Phi-4-MM [5] and Qwen3-Omni [6] demonstrating native audio-text processing. Unlike traditional pipelines that separate ASR from textual analysis, these models process speech directly. Our system using Phi-4-MM to reduce transcription errors and improve inference speed for real-time interaction.

Retrieval-Augmented Generation (RAG) RAG mitigates hallucinations by grounding LLM responses in external factual evidence [7]. However, the effectiveness of RAG relies on semantic retrieval rather than keyword matching. Benchmarks like MTEB [8] highlight the superiority of retrievers such as Qwen3-Embedding for capturing nuance. We integrate these into a RAG framework for course management, enabling the system to resolve complex, domain-specific queries.

3 Approach

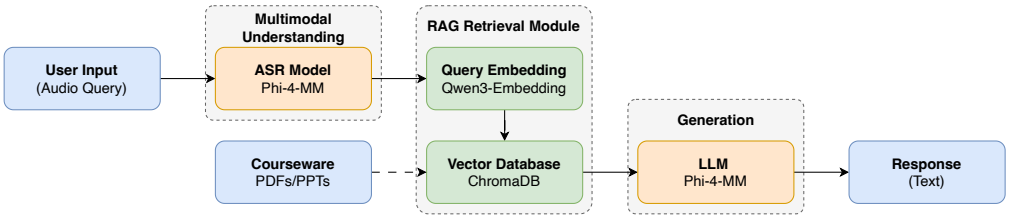

Figure 1: System Architecture of the Speech-RAG Course Assistant. The pipeline integrates multimodal speech understanding with a specialized retrieval module.

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

As shown in Figure 1, we propose a pipeline that transforms raw audio into context-grounded responses. Implementation details are provided in Appendix A

Multimodal Understanding The process begins by capturing the student's voice query. We utilize **Phi-4-MM** to process the audio signal directly, generating high-fidelity transcriptions that preserve the semantic nuance of the question, effectively preparing the input for the retrieval chain.

Retrieval Module The transcribed query is encoded into a dense vector using **Qwen3-Embedding** within the **LangChain**[3] framework. This embedding is used to perform a semantic search against **ChromaDB**[4], a vector database containing indexed course materials (PDFs and PPTs), ensuring the retrieval of precise lecture segments relevant to the inquiry.

Generation Finally, we construct a composite prompt combining the retrieved course context and the user's original query, synthesizing a factually accurate text response.

4 Experiments

Setup & Results Our evaluation is built on a knowledge base constructed by indexing all course PDFs and slides into a vector store. We generated a **Test Set** of multiple-choice questions using **Gemini** (`gemini-2.5-flash`). These queries were synthesized into speech using **Gemini TTS** (`gemini-2.5-flash-preview-tts`), employing random speaker profiles to simulate realistic acoustic diversity and intonation.

We utilize **Accuracy** as the primary metric, verified automatically against ground truth using the xFinder [9] model. We benchmarked four scenarios to isolate the impact of retrieval and modality: **No-RAG** vs. **Speech-RAG** across both **Text** and **Audio** inputs.

Table 1: Comparison across modalities (N=100).

Method	Modality	Acc. (%)
No-RAG	Text	66.00
No-RAG	Audio	58.00
Speech-RAG (Ab.)	Text	91.00
Speech-RAG	**Audio**	**79.00**

Table 2: Representative correction cases.

Case 1: Lecture Location
Query: "Location for Wednesday lectures?"
No-RAG: "C: SDS office." **Ours:** "B: TB 102."

Case 2: Grading Scheme
Query: "Max percentage weight of Final Exam?"
No-RAG: "A: 30 percent." **Ours:** "C: 45 percent."

As shown in Table 1, retrieval is the critical factor for reliability. Incorporating RAG improves accuracy by 25% for text and 21% for audio inputs. Although acoustic processing introduces minor degradation compared to pure text (91% vs. 79%), our end-to-end **Speech-RAG** system significantly outperforms the text-only baseline without RAG (66%). This confirms that grounding the model in course materials effectively neutralizes hallucinations, even amidst the noise of speech understanding.

Case Studies We highlight two cases where general-purpose models fail. As illustrated in Table 2, the No-RAG baseline hallucinates plausible but incorrect logistics. In contrast, our system retrieves specific document chunks (e.g., distinguishing "TB 102" from generic office locations).

5 Conclusion

We presented a Speech-RAG Course Assistant designed to bridge the gap between hands-free interaction and factual precision in educational settings. By synergizing the multimodal capabilities of Phi-4-MM with the dense retrieval of Qwen3-Embedding, we established a robust pipeline that grounds spoken queries directly in course materials. Our evaluation demonstrates a 79% accuracy rate on audio inputs, significantly outperforming the 58% baseline of general-purpose models. Future work will focus on optimizing inference latency for fluid, real-time conversation and scaling the architecture to support multi-course curricula simultaneously.

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

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] 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.

[6] 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.

[7] 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.

[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] 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.

A Implementation Details

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.