

Enhancing Speech-to-Speech Dialogue Modeling with End-to-End Retrieval-Augmented Generation

Pengchao Feng¹, Ziyang Ma¹, Wenxi Chen¹, Yao Li³, Sheng Wang³, Kai Yu¹, Xie Chen^{1,2} *

¹MoE Key Lab of Artificial Intelligence, Jiangsu Key Lab of Language Computing, X-LANCE Lab, School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

²Shanghai Innovation Institute ³Shanghai Aviation Electric Co., Ltd

Abstract

End-to-end speech-to-speech (S2S) dialogue systems have recently garnered increasing research attention for their lower latency and more natural integration of nonverbal cues such as emotion and speaker identity. However, these systems face key challenges, particularly in incorporating external knowledge, a capability commonly addressed by Retrieval-Augmented Generation (RAG) in text-based large language models (LLMs). The core difficulty lies in the modality gap between input speech and retrieved textual knowledge, which hinders effective integration of information. To address this issue, we propose a novel end-to-end RAG framework that directly retrieves relevant textual knowledge from speech queries. Experimental results demonstrate that our method significantly improves the performance of end-to-end S2S dialogue systems while achieving higher retrieval efficiency. Although the overall performance still lags behind the SOTA cascaded models, our framework offers a promising direction for enhancing knowledge integration in end-to-end S2S systems. Our code and dataset are released.

¹

1 Introduction

The release of GPT-4o (Hurst et al., 2024), a unified model capable of comprehending and generating text, speech, and visual modalities within a single neural architecture, represents a significant milestone in the development of Multimodal Large Language Models (MLLMs). MLLMs are defined as large language model-based architectures that can perceive (Dosovitskiy et al., 2020; Chu et al., 2023), reason over (Zhang et al., 2023; Ma et al., 2025), and generate (Zhan et al., 2024; Du et al.,

2023) multimodal information (Liang et al., 2024), with representative models including AnyGPT (Zhan et al., 2024) and Qwen-Omni (Xu et al., 2025). Among various MLLM paradigms, end-to-end speech-to-speech (S2S) dialogue models, such as Mini-Omni (Xie and Wu, 2024), LLaMA-Omni (Fang et al., 2024), and SLAM-Omni (Chen et al., 2024b), constitute an important subclass, aiming to directly respond to speech input with speech output through a single integrated framework. However, due to the nature of the modality gap between input speech and retrieved textual knowledge, these models lack the capability to retrieve external knowledge documents and face limitations in timely knowledge updates and domain-specific information acquisition, hindering their performance in question-answering tasks.

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), which retrieves relevant information from an external source and incorporates it as contextual input to enhance response accuracy, effectively addresses the limitation of text-based LLMs in accessing external knowledge (Gao et al., 2023). However, in end-to-end speech-to-speech dialogue models, directly retrieving textual documents is often ineffective due to the difficulty of aligning input speech with the corresponding textual knowledge. A straightforward solution is to leverage Automatic Speech Recognition (ASR) to convert speech into text and then use the recognized text for retrieval. While this approach can be effective, high-accuracy ASR models, such as Whisper (Radford et al., 2023), introduce additional latency, which poses a critical limitation for real-time vertical spoken assistants.

Therefore, balancing retrieval accuracy and latency remains a critical challenge in building high-performance voice-based retrieval assistants. Given that RAG relies on encoding textual information into a vector space for indexing and retrieval, we propose an end-to-end approach that embeds

* Corresponding Author.

¹The code is available in <https://github.com/the-bird-F/GLM-Voice-RAG>. The speech dataset is available in https://huggingface.co/datasets/the-bird-F/HotpotQA_RGBzh_speech.

speech queries and textual content into a shared representation space, enabling direct speech-to-text retrieval without needing ASR. This approach significantly reduces retrieval latency and streamlines the pipeline.

In our experiments with speech-to-speech dialogue models, the proposed end-to-end RAG approach reduces retrieval latency to one-fourth of that incurred by the ASR-based cascaded RAG models. Specifically, on spoken question-answering tasks, our model achieves an effective improvement in answer accuracy over the baseline without retrieval augmentation (eg, 20% improvement on HotpotQA (Yang et al., 2018)), although performing lower than the cascade models employing ASR-based cascade retrieval. These findings highlight the effectiveness of our approach and establish a promising design paradigm for future research on multimodal RAG.

In summary, our key contributions are as follows:

- We propose an effective end-to-end RAG framework for speech-to-speech dialogue models, which is able to retrieve textual knowledge with input speech directly.
- We conduct a comprehensive evaluation on multilingual benchmarks, which not only demonstrates that the end-to-end retrieval method achieves a 4× speedup in retrieval, but also underscores the critical role of semantic alignment in retrieval accuracy.
- To promote reproducibility and further research, we open-source the code and dataset used in our experiment.

2 Related Work

RAG for Speech-to-Speech Dialogue Models

ReSLM (Wang et al., 2024b) introduces the concept of a speech retriever for the speech-language model SLM (Wang et al., 2023), which leverages retrieved audio context for more grounded responses and lays the foundation for this area.

ReCap (Ghosh et al., 2024) and SRAG (Yang et al., 2024) extend speech retrieval techniques to audio captioning and spoken language understanding, respectively.

SpeechRAG (Min et al., 2024) and SEAL (Sun et al., 2025) are among the first works to explore the use of a shared embedding space for enabling

speech-to-text retrieval. And next WavRAG (Chen et al., 2025) leverages the audio encoding capabilities of Qwen-Audio to generate semantically rich speech embeddings for retrieval.

Our work is contemporaneous with approaches such as SpeechRAG, SEAL, and WavRAG, all driven by the recent surge in spoken dialogue systems and aimed at universal cross-modal RAG. What distinguishes our work is that it not only highlights the trade-offs between ASR-based RAG systems and an end-to-end retrieval system, but also presents multilingual experiments demonstrating superior generalization across languages.

Speech-Text Alignment

Speech-text alignment is a foundational task in speech processing, underpinning key applications such as automatic speech recognition (ASR) and text-to-speech (TTS) synthesis.

Recent models such as SeamlessM4T (Fan et al., 2023) and Whisper (Radford et al., 2023) have achieved accurate and robust performance across diverse languages and acoustic conditions. However, in speech-to-text retrieval tasks, high ASR accuracy does not necessarily lead to better performance, as ASR introduces latency, and full transcriptions are often unnecessary.

SONAR (Duquenne et al., 2023) is a multi-encoder embedding framework that constructs a sentence embedding space from text and uses it as a teacher model to train a speech encoder aligned with the corresponding textual representation. The alignment enables direct retrieval of textual information from speech queries.

3 Method

3.1 System overview

Figure 1 illustrates our proposed end-to-end Retrieval-Augmented Generation system architecture. The system comprises three integrated stages: indexing, retrieval, and generation. Unlike conventional text-based RAG systems, our framework uniquely supports both speech and text modalities across all processing stages.

3.2 Indexing

This stage aims to construct a database from existing text documents.

$$C = \{(\varphi_{\text{text}}(d_i), d_i) \mid D \rightarrow \{d_i\}\} \quad (1)$$

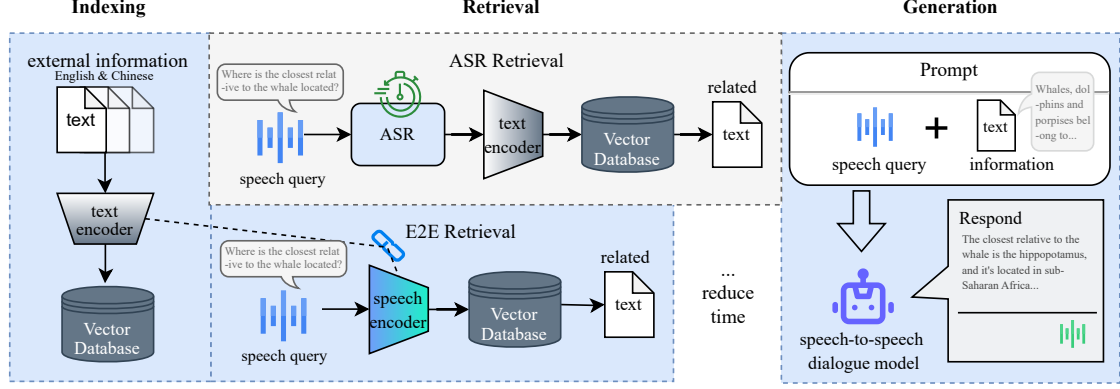


Figure 1: Our proposed framework of end-to-end RAG for speech-to-speech dialogue model. 1) Indexing stage, where text information is divided into chunks and encoded index. 2) Retrieval stage, where we compare ASR-based cascade retrieval, which transcribes speech to text before performing text retrieval, and end-to-end retrieval, which encodes speech queries into the same vector space as texts for direct retrieval. 3) Generation stage, where the end-to-end S2S dialogue model generates responses.

As formula 1, the process involves three key steps: (1) the text documents D are divided into chunks $\{d_i\}$. (2) Each chunk is embedded using the text encoder φ_{text} from a multimodal embedding model, which maps different modalities into a shared vector space. In this work, we adopt SONAR (Duquenne et al., 2023) as the encoder function φ , which comprises modality-specific encoders for text (φ_{text}) and speech (φ_{speech}), respectively. (3) Finally, the encoded indexes and their corresponding text chunks are stored in the database C .

3.3 Retrieval

The retrieval stage aims to extract relevant textual information in response to the spoken query.

$$d_r = \arg \text{top-k sim}(\varphi_{\text{speech}}(q), \varphi_{\text{text}}(d)) \quad (2)$$

As formula 2, we employ the speech encoder φ_{speech} , derived from the same embedding model as φ_{text} , to encode the entire spoken query, thereby obviating the need for speech tokenization. Then the retrieval module computes the similarity between the query embedding and the indexed data chunks, and returns the top- k most similar chunks, which constitute the relevant retrieved information.

By eliminating the need to convert speech to text, the retrieval process reduces the time and enhances overall efficiency.

3.4 Generation

In the final stage, the retrieved information is combined with the original spoken query and fed into the end-to-end speech-to-speech dialogue model. This enables the model to generate responses grounded in factual knowledge, improving its relevance and practicality for real-world applications.

To integrate the user query with the retrieved information and guide the model toward generating factually grounded responses, we use the prompt provided by the langchain (Chase, 2022), which is provided in the appendix.

4 Experiments

4.1 Dataset

HotpotQA (Yang et al., 2018) is a diverse and interpretable question-answering dataset that necessitates multi-hop reasoning. The test-distractor dataset, comprising 7,405 questions and answers, was selected for evaluation. We call Azure’s text-to-speech API ² to synthesize the corresponding speech data.

RGB (Chen et al., 2024a) is an RAG benchmark, available in both English and Chinese, which constructs QA data based on the latest news from 2024. For the experiment, we selected the Chinese test set and used the Azure API to generate the corresponding speech data.

²<https://learn.microsoft.com/en-us/azure/ai-services/speech-service>

Method	Embedding	HotpotQA (en)			RGB (zh)		
		retrieval.t ↓	retrieval.f1 ↑	answer.acc ↑	retrieval.t ↓	retrieval.f1 ↑	answer.acc ↑
w/o RAG	-	-	-	0.27	-	-	0.17
ASR RAG	OpenAI	1.24 s*	0.27	0.48	1.23 s*	0.42	0.68
	BCE	0.41 s	0.25	0.48	0.34 s	0.44	0.69
	M-E5	0.43 s	0.28	0.52	0.33 s	0.47	0.72
E2E RAG	SONAR	0.08 s	0.24	0.43	0.07 s	0.31	0.54
Oracle RAG	OpenAI	-	0.28	0.50	-	0.42	0.69
	BCE	-	0.25	0.48	-	0.41	0.68
	M-E5	-	0.28	0.53	-	0.46	0.73
Facts	-	-	-	0.69	-	-	0.94

Table 1: Performance on two benchmarks: HotpotQA (English) and RGB-zh (Chinese). "w/o RAG" and "Facts" represent retrieval-free and oracle-knowledge settings, respectively. We evaluate ASR RAG and Oracle RAG methods with different embedding models (OpenAI, BCE, M-E5). *: may be affected by the network environment.

4.2 Baseline systems

We construct four system groups for comparison, with their configs provided in the appendix.

- **w/o RAG** This baseline system allows the dialogue model to answer questions directly, without retrieving external knowledge, thereby reflecting the model’s intrinsic knowledge.
- **Facts** This topline system bypasses retrieval and directly accesses the relevant information. It aims to evaluate the comprehension, analysis, and reasoning capabilities of the dialogue model itself.
- **ASR RAG** These systems convert the input speech query into text using an ASR system, and then apply text-based RAG. They represent the most straightforward approach to speech-text retrieval and are designed to contrast the differences between cascading RAG and end-to-end RAG in terms of retrieval efficiency and answer accuracy.
- **Oracle RAG** These systems typically assume that ASR transcriptions are error-free, allowing RAG to be applied directly on the transcribed text without correction. They serve as accuracy benchmarks for text-based RAG to identify issues in implementation or modal alignment.

4.3 Result

The results are shown in Table 1. We examine the effect of the model from two aspects: retrieval efficiency and answer accuracy of the whole dialogue system.

Retrieval Efficiency For each retrieval, we measured the time taken to retrieve information and the similarity of the retrieved information to the ground-truth facts (i.e., the information that can effectively answer the question), which was measured using the F1 score between two text strings. The results show that our end-to-end retrieval system can reduce the retrieval time of the cascade system (ASR RAG) to 1/4 on both the English and Chinese datasets (reducing 0.4s to 0.08s). And the retrieval accuracy of our model is 1-4 percentage points lower than that of the cascade model.

Answer Accuracy For the whole speech-to-speech dialogue system, we use the evaluation metric (Recall, covered Exact Match) provided by the benchmarks to measure the accuracy of the system’s responses. From the results, we can see that our system has a significant improvement over the native dialogue model without RAG (~ 20% improvement on HotpotQA and ~ 43% improvement on RGB-zh); However, it is lower than the model using text directly and the model using high-accuracy ASR retrieval (9% lower than the system using the best embedding on HotpotQA and ~ 15% lower than the system using RGB-zh).

The result shows that in the scenario where response rate is not emphasized, the ASR-based cascade model still needs to be used, which has higher answer accuracy. However, considering some scenarios that require rapid response, such as E-commerce promotion and concert ticket buying, end-to-end RAG is a good trade-off between retrieval efficiency and answer accuracy, which has higher practical value.

4.4 Ablation study

The additional ablation experiment, where the speech encoder was removed and replaced with an ASR retrieval pipeline using Whisper and the SONAR text encoder to perform retrieval, or an alternative speech-text alignment tool CLAP (Elizalde et al., 2023), shows the following results:

Speech Encoder	HotpotQA		RGB	
	re.f1↑	an.acc↑	re.f1↑	an.acc↑
SONAR(ours)	0.24	0.43	0.31	0.54
$\times$	0.25	0.44	0.30	0.56
CLAP	0.14	0.39	0.16	0.42

Table 2: Ablation study evaluating the contribution of SONAR speech encoder by comparing its presence or absence, which replaces it with an ASR pipeline (Whisper and SONAR text encoder, marked $\times$), or CLAP.

As shown in Table 2, when the SONAR text encoder is used for text embeddings, using or removing the SONAR speech encoder yields nearly identical accuracy, which indicates that the accuracy gap between ASR RAG and E2E RAG is not due to the inclusion of the speech encoder, but rather due to the limitations of the SONAR text encoder.

From another perspective, using the SONAR speech encoder with a text decoder for speech recognition tasks (Duquenne et al., 2023) shows that SONAR performs comparably to Whisper v3, suggesting that the SONAR speech encoder introduces minimal additional error.

Additionally, we analyzed failed retrieval cases (Appendix A.4), which show that E2E retrieval using SONAR retrieves fewer semantically relevant segments, particularly when no lexical overlap exists. This supports that the shared embedding space from SONAR lacks strong semantic alignment compared to more advanced encoders like BCE.

Table 2 also demonstrates that replacing SONAR with CLAP leads to a drop in accuracy. The reason is that CLAP is optimized for speech-text alignment but less effective in preserving fine-grained text semantic alignment, as it does not construct a highly semantically coherent embedding space. This result again confirms that **semantic alignment of the shared vector space is a key factor for improving retrieval accuracy**.

And in SONAR training, the text encoder serves as the teacher, and the speech encoder is optimized

via MSE loss to align its output with the corresponding text embedding. Therefore, using more powerful text encoders (e.g., BCE, M-E5, and OpenAI embedding) as the teacher model could theoretically lead to better-trained speech encoders with a stronger semantic vector space, and also enable higher accuracy in E2E RAG systems without sacrificing efficiency.

5 Conclusion

In this work, we highlight the efficiency bottleneck in cross-modal retrieval, spark discussion around the trade-off between latency and accuracy, and contribute an open-sourced, multilingual end-to-end retrieval augmented generation framework to facilitate further research. The proposed method leverages a pretrained embedding model with multiple encoders, allowing both speech and text to be indexed within the same vector space, which enables direct retrieval of relevant textual information from speech queries. The experimental results demonstrate that our method significantly enhances retrieval speed, which is an efficient and scalable framework for end-to-end speech-to-speech dialogue systems. Furthermore, we point out that the key factor in improving retrieval accuracy is enhancing the semantic alignment of the shared vector space, which provides guidance for future research.

Limitations

Despite the promising results of our proposed end-to-end RAG framework for spoken dialogue systems, there are still notable limitations that warrant further investigation.

At the speech-text alignment level, one major limitation is the retrieval accuracy constrained by the suboptimal performance of the SONAR embedding model used in our system. This limitation is especially critical when high-precision semantic understanding is required for information retrieval.

At the RAG level, another significant limitation is that our current approach utilizes only the native RAG model for experimentation. It remains unclear whether adopting more advanced RAG frameworks, such as graph RAG, could further improve the performance of end-to-end speech-to-speech dialogue systems. Future work could investigate the potential benefits of integrating these frameworks to better understand their impact on system performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. U23B2018 and No. 62206171), Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102, and Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2024CSJGG01100).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.
- Harrison Chase. 2022. [LangChain](#).
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. Benchmarking Large Language Models in Retrieval-Augmented Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.
- Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, and 1 others. 2024b. SLAM-Omni: Timbre-controllable voice interaction system with single-stage training. *arXiv preprint arXiv:2412.15649*.
- Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025. WavRAG: Audio-integrated Retrieval Augmented Generation for spoken dialogue models. *arXiv preprint arXiv:2502.14727*.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Zhihao Du, Jiaming Wang, Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, and 1 others. 2023. LauraGPT: Listen, attend, understand, and regenerate audio with GPT. *arXiv preprint arXiv:2310.04673*.
- Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.
- B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang. 2023. CLAP learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Angela Fan, Guillaume Wenzek, and et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.04737*.
- Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-Omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2.
- S. Ghosh, S. Kumar, C. K. R. Evuru, R. Duraiswami, and D. Manocha. 2024. Recap: Retrieval-augmented audio captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1161–1165. IEEE.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. [GPT-4o system card](#). *Preprint*, arXiv:2410.21276.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-Augmented Generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodal large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*.
- Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-CoT: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*.
- D. J. Min, K. Mundnich, A. Lapastora, E. Soltanmohammadi, S. Ronanki, and K. Han. 2024. Speech retrieval-augmented generation without

- automatic speech recognition. *arXiv preprint arXiv:2412.16500*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*. PMLR.
- Chunyu Sun, Bingyu Liu, Zhichao Cui, Anbin Qi, Tian-hao Zhang, Dinghao Zhou, and Lewei Lu. 2025. SEAL: Speech embedding alignment learning for speech large language model with Retrieval-Augmented Generation. *arXiv preprint arXiv:2502.02603*.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- M. Wang, W. Han, I. Shafran, Z. Wu, C. C. Chiu, Y. Cao, and Y. Wu. 2023. SLM: Bridge the thin gap between speech and text foundation models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.
- M. Wang, I. Shafran, H. Soltau, W. Han, Y. Cao, D. Yu, and L. El Shafey. 2024b. Retrieval augmented end-to-end spoken dialog models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12056–12060. IEEE.
- Zhifei Xie and Changqiao Wu. 2024. Mini-Omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*.
- H. Yang, M. Zhang, D. Wei, and J. Guo. 2024. SRAG: Speech retrieval augmented generation for spoken language understanding. In *2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT)*, pages 370–374. IEEE.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.
- I NetEase Youdao. 2023. Bcembedding: Bilingual and crosslingual embedding for rag. *Bcembedding: Bilingual and crosslingual embedding for rag*.
- Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.
- Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, and 1 others. 2024. AnyGPT: Unified multimodal LLM with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *Proc. TMLR*.

A Appendix

All datasets and tools are intended for research use, and our use complies with that purpose.

A.1 Experiment Config

The experiments are conducted on a single NVIDIA A100-PCIE-40GB GPU.

For TTS, we used 'JennyNeural', an American female voice, to generate English speech data, and 'XiaoxiaoNeural', a Mandarin-speaking female voice, for Chinese speech data.

And we adopt three text embedding models as a function φ of our index construction. 1) OpenAI embedding (Achiam et al., 2023) which comes from OpenAI; 2) BCEmbedding base v1 (Youdao, 2023) which comes from YouDao by Wangyi. 3) Multilingual-E5-text-embeddings-large (M-E5) (Wang et al., 2024a) which comes from Microsoft.

Furthermore, we employed the end-to-end speech-to-speech dialogue model GLM-4-Voice (Zeng et al., 2024) as the generator for our case study, and used the top-1 response as the answer. GLM-4-Voice³, developed by Zhipu AI, is an end-to-end speech-to-speech model capable of directly understanding and generating speech in both Chinese and English. It supports real-time voice interaction and allows for the modification of voice attributes based on user commands, making it highly suitable for our experimental needs.

The top 4 relevant chunks were retrieved in all of our experiments.

A.2 Prompt Frame

The main part of the prompt is provided by the langchain (Chase, 2022) retrieval tool, which is proven to be reliable. The assistant prompt is used to prompt the large model to transcribe in streaming mode.

Prompt:

Human: You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.

Question: {speech query}

Context: {retrieved text}

assistant: streaming_transcription

The system prompt adopts the GLM-4-Voice prompt format during instruction fine-tuning to ensure the model can better leverage the responses learned through fine-tuning.

Prompt:

system: The User will provide you with a speech instruction. Do it step by step. First, think about the instruction and respond in an interleaved manner, with 13 text tokens followed by 26 audio tokens.

A.3 ASR

We adopt Whisper-large-v3 (Radford et al., 2023) as the ASR model.

The model, when transcribing the speech generated by the Azure Speech Service, introduces a certain level of error, with a word error rate (WER) of 13.0% on English data and a character error rate (CER) of 9.5% on Chinese data. For reference, this version of the Whisper model reports a WER of 9.3% on English and a CER of 12.8% on Mandarin Chinese based on the Common Voice 15 dataset (Ardila et al., 2019) when evaluated independently (Radford et al., 2023).

A.4 Experiment Cases

Success cases

Question:

What science fantasy young adult series, told in first person, has a set of companion books narrating the stories of enslaved worlds and alien species?

Ground Truth:

Animorphs

Useful Information:

'The Hork-Bajir Chronicles is the second companion book to the "Animorphs" series, written by K. A. Applegate.', 'With respect to continuity within the series, it takes place before book 23, "The Pretender", although the events told in the story occur between the time of "The Ellimist Chronicles" and "The Andalite Chronicles".', 'The book is introduced by Tobias, who flies to the valley of the free Hork-Bajir, where Jara Hamee tells him the story of how the Yeerks enslaved the Hork-Bajir, and how Aldrea, an Andalite, and her companion, Dak Hamee, a Hork-Bajir, tried to save their world from the invasion.', 'Animorphs is a science

³<https://github.com/THUDM/GLM-4-Voice>

fantasy series of young adult books written by Katherine Applegate and her husband Michael Grant, writing together under the name K. A. Applegate, and published by Scholastic.’
’ It is told in first person, with all six main characters taking turns narrating the books through their own perspectives.’

Prediction:

It sounds like you’re describing ‘The Animagus’ series. It’s a science fantasy young adult book series where the main characters can transform into animals. The series explores themes like friendship, family, and fighting for justice. If you’re looking for more suggestions, I’d be happy to help!

Result: True

Question:

Are the Laleli Mosque and Esma Sultan Mansion located in the same neighborhood?

Ground Truth:

No

Useful Information:

’The Laleli Mosque (Turkish: "Laleli Camii, or Tulip Mosque") is an 18th-century Ottoman imperial mosque located in Laleli, Fatih, Istanbul, Turkey.’, ’The Esma Sultan Mansion (Turkish: "Esma Sultan Yalısı"), a historical yalı (English: waterside mansion) located at Bosphorus in Ortaköy neighborhood of Istanbul, Turkey and named after its original owner Esma Sultan, is used today as a cultural center after being redeveloped.’

Prediction:

Laleli Mosque and Esma Sultan Mansion are not in the same neighborhood. Laleli Mosque is in the Fatih district, while Esma Sultan Mansion is in the Ortaköy neighborhood, which is part of the Beyoğlu district. They’re both in Istanbul, but they’re quite a distance apart.

Result: True

Failure cases

Question :

The director of the romantic comedy "Big Stone Gap" is based in what New York city?

Ground Truth:

Greenwich Village

Useful Information:

’Big Stone Gap is a 2014 American drama romantic comedy film written and directed by Adriana Trigiani and produced by Donna Gigliotti for Altar Identity Studios, a subsidiary of Media Society.’, ’Adriana Trigiani is an Italian American best-selling author of sixteen books, television writer, film director, and entrepreneur based in Greenwich Village, New York City.’

Prediction:

The director of "Big Stone Gap" is likely not based in New York City, as the film is set in a small town in the United States. However, if you are referring to a specific character or context within the film, please provide more details for me to better assist you.

Result: False

Question:

What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

Ground Truth:

Chief of Protocol

Useful Information:

’Kiss and Tell is a 1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer.’, "Shirley Temple Black (April 23, 1928 – February 10, 2014) was an American actress, singer, dancer, business-woman, and diplomat who was Hollywood’s number one box-office draw as a child actress from 1935 to 1938.", ’ As an adult, she was named United States ambassador to Ghana and to Czechoslovakia and also served as Chief of Protocol of the United States.’

Prediction:

In "Kiss and Tell," Corliss Archer was portrayed by Shirley Temple. However, I don’t know which government position she held in real life or if she held any at all. Temple was best known for her film career, and she never held a government position.

Result: False