GISP@HEU'S SUBMISSION FOR DCASE 2026 TASK 5: A LORA-SFT FINE-TUNED AUDIO-DEPENDENT QUESTION ANSWERING SYSTEM

Technical Report

Feiyang Xiao[1], *Qiaoxi Zhu*[2], *and Jian Guan*[1*]

[1]College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]University of Technology Sydney, Ultimo, Australia

ABSTRACT

This technical report presents our submission for DCASE 2026 Task 5: Audio-Dependent Question Answering (ADQA), which aims to evaluate a model's ability to answer natural language questions based on information contained in audio signals. In our submission, we develop a system based on the pretrained audio-language model (i.e., Fun-Audio-Chat) to the ADQA task through parameter-efficient supervised fine-tuning using Low-Rank Adaptation (LoRA). The resulting system exploits audio content and question semantics to generate accurate answers for audio-dependent question answering.

***Index Terms*—** Audio-dependent question answer, audio-language model, low-rank adaptation

1. INTRODUCTION

Audio-Dependent Question Answering (ADQA), focuses on evaluating a model's ability to answer natural language questions by exploiting information present in audio signals [1]. Unlike conventional text-based question answering [2, 3], ADQA requires models to extract and integrate multiple levels of information from audio, including acoustic events, spoken content, speaker characteristics, environmental context, and other audio cues, in order to produce correct answers [4].

In this work, we present our submission to DCASE 2026 Task 5. Our system is based on Fun-Audio-Chat [5], a general-purpose audio-language model capable of processing diverse audio inputs and generating text responses. To adapt the pretrained model to the ADQA task, we employ parameter-efficient supervised fine-tuning using Low-Rank Adaptation (LoRA) [6]. The resulting system leverages the strong audio understanding capabilities of Fun-Audio-Chat while requiring only a small number of trainable parameters. The proposed approach provides an efficient and scalable solution for audio-dependent question answering.

2. SUBMISSION SYSTEM

Our submission system is built upon Fun-Audio-Chat [5] as the backbone audio-language model. Given an audio signal and a corresponding natural language question, the model first extracts acoustic representations through its pretrained audio encoder. The encoded audio features are then projected into the language model space and jointly processed with the textual question. This multimodal representation enables the model to capture both acoustic and linguistic information, allowing it to generate an answer conditioned on the audio content. Since Fun-Audio-Chat has been pretrained on diverse audio-language tasks, it provides a strong foundation for audio understanding and audio-grounded reasoning.

To further improve performance on ADQA, we apply supervised fine-tuning (SFT) with Low-Rank Adaptation (LoRA). Instead of updating all model parameters, LoRA inserts trainable low-rank matrices into Transformer layers while keeping the majority of pretrained parameters frozen. This parameter-efficient strategy significantly reduces computational cost and memory consumption during training. The model is fine-tuned using the official ADQA training data, i.e., AudioMCQ-StrongAC-GeminiCoT [1], where each training instance consists of an audio clip, a question, and its corresponding answer. During optimization, the model learns to align audio evidence with question semantics and generate the target answer through an auto-regressive objective. The final system combines the general audio understanding capability of Fun-Audio-Chat with task-specific knowledge acquired through LoRA-based supervised fine-tuning, resulting in an effective solution for audio-dependent question answering.

3. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed system, we conducted experiments on the official development set of DCASE 2026 Task 5. We first reproduced the original Fun-Audio-Chat model as our baseline and evaluated it under the same inference setting. The proposed system was obtained by applying LoRA-based supervised fine-tuning to the baseline model using the official ADQA training data. Results are shown in Table 1.

Table 1: Performance comparison between our systems and the baseline on the official development dataset.

Method	Top-1 Acc (%)
Baseline	52.96
Our System	**53.08**

Experimental results show that the proposed system consistently outperforms the reproduced Fun-Audio-Chat baseline on the development set. The performance improvement indicates that task-specific supervised fine-tuning effectively enhances the model's ability to associate audio evidence with question semantics and generate more accurate answers. These results demonstrate the

*Corresponding author.

effectiveness of the proposed adaptation strategy for the Audio-Dependent Question Answering task.

4. CONCLUSION

This technical report presented our submission to DCASE 2026 Task 5: Audio-Dependent Question Answering. Our system is built upon Fun-Audio-Chat and adapted to the ADQA task through parameter-efficient LoRA-based supervised fine-tuning. Experiments on the official development set demonstrate that the proposed system outperforms the reproduced Fun-Audio-Chat baseline, confirming the effectiveness of task-specific adaptation. Future work will explore more advanced audio reasoning strategies and larger-scale audio-language instruction tuning to further improve performance on audio-dependent question answering.

5. ACKNOWLEDGMENT

This work was supported by the Heilongjiang Provincial Natural Science Foundation of China under Grant No. BS2025F009.

6. REFERENCES

[1] H. He, X. Du, R. Sun, Z. Dai, Y. Xiao, M. Yang, J. Zhou, X. Li, Z. Liu, Z. Liang, C. Wu, Q. He, T. Lee, X. Chen, W.-L. Zheng, W. Wang, M. Plumbley, J. Liu, and Q. Kong, "Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2026.

[2] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "MMAU: A massive multi-task audio understanding and reasoning benchmark," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2025.

[3] Z. Ma, Y. Ma, Y. Zhu, C. Yang, Y.-W. Chao, R. Xu, W. Chen, Y. Chen, Z. Chen, J. Cong, K. Li, K. Li, S. Li, X. Li, X. Li, Z. Lian, Y. Liang, M. Liu, Z. Niu, T. Wang, W. Yuping, Y. Wang, Y. Wu, G. Yang, J. Yu, R. Yuan, Z. Zheng, Z. Zhou, H. Zhu, W. Xue, E. Benetos, K. Yu, E.-S. Chng, and X. Chen, "MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, vol. 38. Curran Associates, Inc., 2025.

[4] D. Wang, J. Li, J. Wu, D. Yang, X. Chen, T. Zhang, and H. Meng, "MMSU: A massive multi-task spoken language understanding and reasoning benchmark," *arXiv preprint arXiv:2506.04779*, 2025.

[5] T. F. Team, Q. Chen, L. Cheng, C. Deng, X. Li, J. Liu, C.-H. Tan, W. Wang, J. Xu, J. Ye, *et al.*, "Fun-Audio-Chat technical report," *arXiv preprint arXiv:2512.20156*, 2025.

[6] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2022.