PARAMETER-EFFICIENT FINE-TUNING OF AUDIO FLAMINGO 2 WITH LORA FOR THE DCASE2025 AUDIO QUESTION ANSWERING CHALLENGE

Technical Report

HaeChun Chung

Republic of Korea
marcomx1@naver.com

ABSTRACT

Audio Question Answering (AQA) presents a significant challenge, demanding models capable of complex reasoning over extensive audio sequences. In this research, we boost the performance of Audio Flamingo 2 (AF2), a compact yet powerful audio-language model, by employing parameter-efficient Low-Rank Adaptation (LoRA). We apply targeted data augmentation strategies for multiple-choice QA and fine-tune the model using the DCASE2025 Challenge Task 5 dataset. Our top-performing model, utilizing LoRA with a rank of 8, achieves a remarkable 69.67% accuracy. This substantially outperforms all established baselines, including the strong Gemini-2.0-Flash (52.5%). These results highlight the effectiveness and practical value of lightweight adaptation approach, especially when operating under constrained computational resources.

Index Terms— Audio Question Answering, Audio-Language Model, Audio Flamingo 2, LoRA

1. INTRODUCTION

Audio Question Answering (AQA) is an emerging task in multimodal artificial intelligence that aims to develop systems capable of understanding complex audio content and responding accurately to natural language questions. Unlike conventional audio classification, AQA requires high-level semantic comprehension, temporal reasoning, and contextual understanding across long audio sequences. The DCASE2025 Challenge Task 5 [1, 2] provides a structured benchmark for AQA through three distinct domains: Bioacoustics QA, Temporal Soundscapes QA, and Complex QA, each emphasizing different aspects of audio understanding.

Several approaches have been proposed to tackle the challenges of AQA, including baseline models provided in the DCASE2025 Challenge such as Qwen2-Audio-7B [3], Audio Flamingo 2 [4], and Gemini-2.0-Flash [5]. Among available foundation models, we selected Audio Flamingo 2 (AF2) as our base architecture due to its balanced trade-off between model size and performance. Unlike much larger models, AF2 features a compact 3-billion-parameter language model paired with a powerful CLAP audio encoder [6] and a Flamingo-style cross-attention mechanism [7, 8]. This design balances model capacity and efficiency, enabling strong reasoning over the audio without requiring extensive computational resources.

To enhance AF2's performance on the AQA task, we employed Low-Rank Adaptation (LoRA) [9], a parameter-efficient fine-tuning strategy. LoRA was chosen for several reasons: it allows effective adaptation of large pre-trained models with minimal computational cost, mitigates the risk of catastrophic forgetting by keeping the original model weights intact, and enables flexible experimentation under resource constraints.

Table 1: Performance comparison on evaluation set

Model	Accuracy (%)
Baseline	
Qwen2-Audio-7B	45.0
AudioFlamingo2	45.7
Gemini-2.0-Flash	52.5
Ours	
AudioFlamingo2 + LoRA (rank 8)	**69.67**
AudioFlamingo2 + LoRA (rank 16)	68.91
AudioFlamingo2 + LoRA (rank 32)	68.26

Through systematic experimentation with different LoRA rank configurations applied to the attention mechanisms across all major components of AF2, our approach achieved remarkable performance improvements. The best configuration with LoRA rank 8 attained 69.67% accuracy on the development set, representing a substantial 24% improvement over the baseline AF2 model and surpassing the strongest existing baseline, Gemini-2.0-Flash, by over 17%. These results demonstrate both the effectiveness of our parameter-efficient fine-tuning approach and the exceptional potential of AF2 as a foundation model for AQA tasks.

2. METHOD

We applied LoRA to the query, key, and value projection layers across the three major components of AF2: the CLAP encoder, the Audio Transformer, and the Language Model. This ensures that the adaptation effectively influences the attention mechanisms critical to audio-text reasoning. By adjusting attention weights without altering the core model parameters, we achieved efficient adaptation while preserving the strengths of the original architecture.

To explore the balance between fine-tuning efficiency and model performance, we conducted experiments using three different LoRA rank settings: 8, 16, and 32. The rank-8 configuration introduced approximately 3.38 million trainable parameters, providing a highly parameter-efficient solution. Increasing the rank to 16 doubled the number of trainable parameters to 6.75 million, allowing greater adaptation capacity while still maintaining a lightweight footprint. The rank-32 configuration expanded the trainable set to 13.5 million parameters, offering the highest flexibility among the

three while remaining significantly more efficient than full model fine-tuning.

Because the AQA task is structured as multiple-choice question answering, input formatting plays a crucial role in model performance. A key component of our methodology involved data augmentation techniques specifically tailored to this format. During training, we randomly shuffled the order of answer choices and adjusted the corresponding labels accordingly, while ensuring that the original content remained semantically intact. This strategy was designed to prevent the model from developing positional bias and to promote robustness to variations in choice ordering.

In addition, we applied a label randomization technique to further improve the model's generalization ability. Each training example was randomly assigned one of six labeling formats: uppercase letters with periods (A., B., C., D., E., F.), uppercase letters in parentheses ((A), (B), (C), (D), (E), (F)), lowercase letters with periods (a., b., c., d., e., f.), lowercase letters in parentheses ((a), (b), (c), (d), (e), (f)), numbers with periods (1., 2., 3., 4., 5., 6.), and numbers in parentheses ((1), (2), (3), (4), (5), (6)). By exposing the model to varied formats, we minimized its reliance on specific label conventions and improved its ability to handle diverse question presentations. For evaluation, we standardized the format by consistently using uppercase letters in parentheses ((A), (B), (C), (D), (E), (F)) without shuffling the order of choices. This allowed for fair and consistent performance measurement while leveraging the format robustness gained during training.

3. EXPEIMENTS

We used only the development dataset provided by the challenge organizers for model training and evaluated our models on the official evaluation dataset. To assess performance, we compared our results against the baseline models released with the challenge. Among them, Qwen2-Audio-7B and Audio Flamingo 2 each achieved approximately 45% accuracy, while Gemini-2.0-Flash recorded the highest baseline performance at 52.5%. Our fine-tuned models outperformed all baseline models by a notable margin. In particular, the configuration using LoRA with rank 8 achieved 69.67% accuracy on the evaluation set—representing a 24% improvement over the AF2 baseline and a 17% gain over the strongest baseline, Gemini-2.0-Flash. These results highlight the effectiveness of our parameter-efficient fine-tuning approach.

Interestingly, increasing the LoRA rank beyond 8 did not yield additional performance improvements. The model with rank 16 achieved 68.91% accuracy, while rank 32 showed a slight decrease to 68.26%. These results offer meaningful insights into the balance between adaptation capacity and performance in audio question answering tasks. We speculate that this saturation effect stems from the limited size of the training data relative to the number of trainable parameters. While higher-rank LoRA configurations inherently offer greater expressive power and more capacity for adaptation, they may also introduce a higher risk of overfitting when the training dataset is not sufficiently large or diverse to support such extensive parameter tuning. This suggests that, under constrained data conditions, lightweight adaptation methods like LoRA with a carefully selected, lower rank can be a more robust and effective approach than larger-scale tuning, which might simply memorize noise or spurious correlations in the data.

4. CONCLUSION

This work effectively demonstrates the significant potential of parameter-efficient fine-tuning techniques for enhancing audio-language models in audio question answering (AQA) task. Through systematic application of LoRA to the Audio Flamingo 2 (AF2) model, we achieved a substantial performance improvement, reaching 69.67% accuracy—significantly outperforming all baseline models. This shows AF2's robust capability in handling diverse audio question answering scenarios. Despite limited computational resources and the inability to explore larger models or conduct extensive hyper-parameter tuning, our approach represented strong practical value, showing that even lightweight adaptations can yield state-of-the-art results. Notably, the observation that lower-rank configurations outperformed higher ones suggests that constrained parameterization can act as a form of regularization, promoting generalization in data-limited settings. However, greater performance gains may be achievable through heavier fine-tuning on larger datasets, which we leave as future work.

5. REFERENCES

[1] http://dcase.community/challenge2025/.

[2] C.-H. H. Yang, S. Ghosh, Q. Wang, J. Kim, H. Hong, S. Kumar, G. Zhong, Z. Kong, S. Sakshi, V. Lokegaonkar, *et al.*, "Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge," *arXiv preprint arXiv:2505.07365*, 2025.

[3] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[4] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, "Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities," *arXiv preprint arXiv:2503.03983*, 2025.

[5] https://deepmind.google/models/gemini/flash/.

[6] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[8] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," *arXiv preprint arXiv:2402.01831*, 2024.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.