CURRICULUM LEARNING FOR AUDIO-DEPENDENT QUESTION ANSWERING: TECHNICAL REPORT FOR DCASE 2026 TASK 5

Technical Report

Qixuan Huang, Yizhi Pan, Xiajie Zhou, Rui Li, Masashi Unoki

Graduate School of Advanced Science and Technology,
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
{qixuan, panyizhi, xiajie, rui.li, unoki}@jaist.ac.jp

ABSTRACT

This technical report describes our submission to DCASE 2026 Task 5: Audio-Dependent Question Answering. Our system is based on Fun-Audio-Chat-8B and focuses on curriculum learning during supervised fine-tuning. Motivated by the observation that the training set is generally easier than the evaluation set, we investigate how to maximize knowledge acquisition from limited training data through an appropriate training order. To construct the curriculum, we estimate question difficulty using the length of Chain-of-Thought (CoT) reasoning annotations and sort training samples from easy to hard. The resulting curriculum is used to fine-tune the model with LoRA adaptation. Experimental results on the development set demonstrate that easy-to-hard curriculum learning achieves the best performance, reaching 64.84% accuracy and outperforming both random ordering and reverse curriculum training. These findings suggest that difficulty-based sample ordering is an effective strategy for improving audio reasoning performance in Audio Question Answering (AQA).

***Index Terms*—** Audio Multiple-Choice Question Answering, Curriculum Learning, Auditory Large Language Models, Audio Reasoning

1. INTRODUCTION

Audio Question Answering (AQA) aims to answer questions whose solutions depend on information contained in audio recordings. Compared with traditional audio classification tasks, AQA requires models not only to recognize acoustic events and sound attributes, but also to perform higher-level reasoning over temporal relationships, contextual information, and complex auditory scenes. As a result, AQA has become an important benchmark for evaluating the audio understanding and reasoning capabilities of Auditory Large Language Models (ALLMs) [1, 2].

Recent advances in ALLMs, such as Qwen2-Audio [3], Audio Flamingo [4], Kimi-Audio [5], and Step-Audio 2 [6], have significantly improved performance on a variety of audio understanding tasks. Nevertheless, AQA remains challenging due to the diversity of question types and reasoning requirements, spanning sound event recognition, temporal reasoning, counting, causal inference, commonsense reasoning, and combinations thereof. Consequently, the difficulty of training samples can vary substantially across the dataset.

Improving performance commonly involves scaling model size, introducing additional training data, or employing more sophisticated training objectives. However, such approaches are often computationally expensive and may not always be feasible. In our preliminary analysis, as shown in Table 1, we observe that the training set is generally easier than the evaluation set, suggesting a difficulty mismatch between training and inference. Under this setting, an important question arises: how can a model acquire as much knowledge as possible from the available training data and better prepare for more challenging evaluation samples?

Curriculum learning [10], first introduced by Bengio et al., proposes organizing training samples from easy to hard, inspired by how humans typically learn simpler concepts before tackling more complex ones. This strategy improves optimization stability and generalization, and has since gained renewed attention in the era of large foundation models [11]. In natural language processing, curriculum-based training has been shown to improve reasoning performance through difficulty-based sample scheduling [12]. Similar benefits have been demonstrated in vision-language models, where progressive training helps models acquire multimodal reasoning capabilities [13]. However, its effectiveness for AQA remains largely unexplored.

In this technical report, we investigate curriculum learning for DCASE 2026 Task 5: Audio-Dependent Question Answering. We estimate sample difficulty using the length of CoT [14] reasoning annotations, where longer reasoning chains are assumed to reflect more complex reasoning processes. Training samples are sorted from easy to hard and used to fine-tune a Fun-Audio-Chat-8B model. We further compare the proposed curriculum with random ordering and reverse curriculum training to analyze the impact of training order on performance.

The main contributions of this work are as follows:

- We investigate the effect of training order on Audio-Dependent Question Answering.
- We demonstrate that easy-to-hard curriculum learning consistently outperforms both random ordering and reverse curriculum training on the DCASE 2026 Task 5 development set.

2. DATASET

We conduct all experiments using the official training and development sets provided by DCASE 2026 Task 5: Audio-Dependent Question Answering. The task is closely related to recent au-

Table 1: Performance (%) of representative ALLMs on the DCASE 2026 Task 5 training set without task-specific fine-tuning across different question categories.

Model	Size	Music (%)	Sound (%)	Speech (%)	Temporal (%)	Overall (%)
Kimi-Audio [5]	7B	34.05	35.28	44.67	30.02	39.92
Step-Audio 2 mini [6]	7B	83.17	**84.99**	83.82	55.66	82.46
MiMo-Audio [7]	7B	**88.40**	83.80	**86.16**	47.26	**83.81**
Qwen3-Omni [8]	30B	–	–	–	–	–
Fun-Audio-Chat [9]	8B	66.42	65.54	83.34	**62.86**	75.06

dio understanding and reasoning benchmarks such as MMAU [1], MMAR [2], and shares the motivation of recent work on quantifying and improving the actual contribution of audio to model predictions in ALLMs [15].

3. PROPOSED METHOD

3.1. Overview

Our proposed framework aims to improve Audio-Dependent Question Answering through curriculum learning. Given a training set consisting of audio recordings, questions, candidate choices, correct answer, and CoT annotations, we first estimate the difficulty of each training sample. The samples are then sorted according to their estimated difficulty and presented to the model from easy to hard during supervised fine-tuning.

3.2. Difficulty Estimation

A key component of curriculum learning is the estimation of sample difficulty. Since the official training set provides CoT reasoning annotations, we use the length of the reasoning process as a proxy for question difficulty.

Specifically, given the i-th training sample x_i with corresponding CoT annotation r_i, the difficulty score is defined as

$$D(x_i) = |r_i|, \tag{1}$$

where $|r_i|$ denotes the number of tokens in the CoT.

The intuition behind this design is that questions requiring longer reasoning chains generally involve more complex reasoning processes and therefore represent more challenging training examples.

3.3. Curriculum Learning

After estimating the difficulty score of each training sample, we first group all samples according to their question types. Let

$$\mathcal{G} = \{G_1, G_2, \ldots, G_K\}, \tag{2}$$

where K denotes the number of question categories and G_k represents the set of samples belonging to the k-th question type.

For each group G_k, samples are sorted according to their difficulty scores in ascending order:

$$D(x_1^{(k)}) \leq D(x_2^{(k)}) \leq \cdots \leq D(x_{N_k}^{(k)}), \tag{3}$$

where N_k denotes the number of samples in group G_k.

After sorting, samples from different groups are merged while preserving the easy-to-hard ordering within each question type. The resulting sequence forms a type-based curriculum for supervised fine-tuning.

The motivation behind this design is that different question types often exhibit distinct reasoning patterns and CoT length distributions. Directly applying a global difficulty ranking may cause certain question categories to dominate specific stages of training. By performing curriculum learning within each question type, the model can progressively learn more complex reasoning patterns while maintaining balanced exposure to diverse question categories.

4. EXPERIMENT SETUP

We adopt Fun-Audio-Chat-8B as the backbone model and perform Supervised Fine-Tuning (SFT) using the LLaMA-Factory framework [16]. The model is loaded in 4-bit quantization and trained using LoRA adaptation. The LoRA rank and alpha are set to 16 and 32, respectively. Training is conducted for one epoch with a per-device batch size of 4 and a gradient accumulation step of 4, resulting in an effective batch size of 16. We use the AdamW optimizer with a learning rate of 2×10^{-4}. All experiments, including random ordering, reverse curriculum, and the proposed easy-to-hard curriculum, are trained under identical settings to ensure a fair comparison, all experiments are conducted using a single training epoch. The implementation code and experimental scripts are publicly available at: `https://github.com/OrgHuang/DCASE_2026_Task5_Huang_JAIST.git`.

5. RESULTS

5.1. Main Results

Table 2 presents the performance of different audio-language models on the DCASE 2026 Task 5 development set. Compared with the original Fun-Audio-Chat-8B baseline (56.81%), supervised fine-tuning improves the performance to 62.87%, demonstrating the effectiveness of task-specific adaptation.

Applying the proposed curriculum learning strategy further improves the accuracy to 64.84%, achieving the best overall performance among all evaluated systems. Compared with randomly ordered training samples, the proposed curriculum provides an absolute improvement of 1.97%.

5.2. Effect of Curriculum Learning

Table 3 investigates the influence of training sample ordering. To reduce the effect of randomness, we evaluate four different random seeds and report their average performance.

The average accuracy of random ordering is 62.87%, while the proposed easy-to-hard curriculum achieves 64.84%. Notably, the curriculum outperforms all random ordering runs, indicating that

Table 2: Performance (%) on the DCASE 2026 Task 5 development set.

Method	Model	Accuracy (%)
Baseline	Step-Audio 2 mini	50.53
SFT (Random Ordering)	Step-Audio 2 mini	51.65
Baseline	MiMo-Audio-7B	54.57
SFT (Random Ordering)	MiMo-Audio-7B	55.07
Baseline	Fun-Audio-Chat-8B	56.81
SFT (Random Ordering)	Fun-Audio-Chat-8B	62.87
CoT Curriculum (Easy → Hard)	Fun-Audio-Chat-8B	**64.84**
CoT Curriculum (Hard → Easy)	Fun-Audio-Chat-8B	62.85

Table 3: Effect of curriculum learning compared with random sample ordering.

Training Strategy	Accuracy (%)
Random Ordering (Seed 42)	63.47
Random Ordering (Seed 123)	63.53
Random Ordering (Seed 456)	61.73
Random Ordering (Seed 789)	62.73
Random Ordering (Average)	62.87
CoT Curriculum (Easy → Hard)	**64.84**

the observed improvement is unlikely to be caused by random fluctuations during training. These results suggest that organizing training samples according to difficulty allows the model to gradually acquire more complex reasoning capabilities and improves knowledge acquisition from the available training data.

6. CONCLUSION

This technical report presented our submission to DCASE 2026 Task 5: Audio-Dependent Question Answering. We proposed a curriculum learning strategy that organizes training samples from easy to hard based on CoT annotation length, and demonstrated its consistent improvement over both random ordering and reverse curriculum training. While the absolute gain may appear modest, it is worth noting that the backbone model already achieves 75.06% zero-shot accuracy on the training set, leaving limited room for improvement under the current data regime. We hypothesize that difficulty-based curriculum learning may yield more pronounced benefits when applied to larger-scale and more balanced datasets, where the difficulty gap between training samples is more pronounced.

7. ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI (25H01139).

8. REFERENCES

[1] 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 *International Conference on Learning Representations*, vol. 2025, 2025, pp. 84 929–84 964.

[2] Z. Ma, Y. Ma, Y. Zhu, C. Yang, Y.-W. Chao, R. Xu, W. Chen, Y. Chen, Z. Chen, J. Cong, *et al.*, "Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix," *Advances in Neural Information Processing Systems*, vol. 38, 2026.

[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] D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang, *et al.*, "Kimi-audio technical report," *arXiv preprint arXiv:2504.18425*, 2025.

[6] B. Wu, C. Yan, C. Hu, C. Yi, C. Feng, F. Tian, F. Shen, G. Yu, H. Zhang, J. Li, *et al.*, "Step-audio 2 technical report," *arXiv preprint arXiv:2507.16632*, 2025.

[7] D. Zhang, G. Wang, J. Xue, K. Fang, L. Zhao, R. Ma, S. Ren, S. Liu, T. Guo, W. Zhuang, *et al.*, "Mimo-audio: Audio language models are few-shot learners," *arXiv preprint arXiv:2512.23808*, 2025.

[8] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, *et al.*, "Qwen3-omni technical report," *arXiv preprint arXiv:2509.17765*, 2025.

[9] 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.

[10] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[11] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum learning: A survey," *International Journal of Computer Vision*, vol. 130, no. 6, pp. 1526–1565, 2022.

[12] B. Xu, L. Zhang, Z. Mao, Q. Wang, H. Xie, and Y. Zhang, "Curriculum learning for natural language understanding," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 6095–6104.

[13] O. Thawakar, D. Dissanayake, K. P. More, R. Thawkar, A. Heakl, N. Ahsan, Y. Li, I. Z. M. Zumri, J. Lahoud, R. M. Anwer, *et al.*, "Llamav-o1: Rethinking step-by-step visual reasoning in llms," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 24 290–24 315.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[15] H. He, X. Du, R. Sun, Z. Dai, Y. Xiao, M. Yang, J. Zhou, X. Li, Z. Liu, Z. Liang, *et al.*, "Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models," *arXiv preprint arXiv:2509.21060*, 2025.

[16] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, and Z. Luo, "Llamafactory: Unified efficient fine-tuning of 100+ language models," in *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, 2024, pp. 400–410.