# AUDIO QUESTION ANSWERING AT THE DCASE 2026 CHALLENGE

Technical Report

*Haoran Xu*

Huazhong University of Science and Technology
School of Computer Science and Technology
Wuhan, China
m202574318@hust.edu.cn

*Rui Zhang*

Huazhong University of Science and Technology
School of Computer Science and Technology
Wuhan, China
rui.zhang@ieee.org

## ABSTRACT

In this technical report, we present our system for Task 5, "Audio Question Answering," of DCASE 2026. For this task, we propose a skill-aware audio question answering framework. We classify audio question-answering samples into ten skill categories, including speech content, speaker identity, sentiment and prosody, temporal reasoning, sound event understanding, music style, speech perception, acoustic scene understanding, multi-source reasoning, and other questions. Based on this classification scheme, we annotated a large-scale audio question-answering dataset named SkillBench-AQA, which contains 524,737 question-answer pairs, 365,670 audio clips, 10 skills, and 34 sub-skills. Using SkillBench-AQA and the DCASE 2026 Task 5 data, we developed a system based on MOSS-Audio-4B. This system comprises full-dataset supervised fine-tuning, skill-specific adapters, and a skill router.

Experimental results demonstrate that this approach significantly improves the accuracy of multiple-choice audio question-answering systems. MOSS-Audio-4B achieved a % accuracy on the DCASE 2026 Task 5 development set.

***Index Terms*—** DCASE 2026, Audio Question Answering, Audio Language Model, Skill Router, Supervised Fine-Tuning

## 1. INTRODUCTION

Audio Question Answering (AQA) is a challenging task that requires models to understand audio signals and answer natural language questions about the audio content [1, 2]. Compared with traditional audio classification or captioning tasks, AQA requires the model to perform more fine-grained perception and reasoning over speech, sound events, music, speakers, temporal relations, and complex acoustic scenes.

Recent large audio language models such as Qwen-Audio [3], Qwen2-Audio [4], Kimi-Audio [5], Audio Flamingo 2 [6], and MOSS-Audio [7] have achieved strong performance on audio understanding benchmarks through instruction tuning, supervised fine-tuning (SFT), and reinforcement learning.

In this technical report, we focus on a skill-aware view of audio question answering. Previous analysis has studied whether a model actually listens to the audio input. We further ask: after the model has listened to the audio, does it understand the audio content, and to what extent? To answer this question, we divide AQA questions into ten skill categories and evaluate the model under each skill.

Our preliminary experiments show that the model performs very differently across skills. It can handle some skills relatively well, while performing poorly on others. Moreover, when we apply full-data supervised fine-tuning, the overall accuracy changes only slightly. A detailed skill-level analysis shows that this is because some skills improve while others degrade, making the overall improvement limited.

Based on these observations, we build a skill-aware system for DCASE 2026 Task 5. The system is based on MOSS-Audio-4B and includes full-data SFT, skill-specific adapters, and a skill router. We also construct SkillBench-AQA, a large-scale skill-annotated AQA dataset for training and analysis.

## 2. SKILL TAXONOMY

To perform skill-level analysis, we define ten audio understanding skills for AQA. The skill taxonomy is shown in Table 1.

In addition to the ten main skills, we further define 34 sub-skills for more detailed annotation. These sub-skills include speech transcription, speech semantics, dialogue context, speaker count, speaker identity, speaker demographics, voice quality, emotion recognition, prosody tone, vocal reaction, event presence, event sequence, event counting, source identification, action recognition, environmental sound, mechanical vehicle, animal sound, human nonverbal sound, object material sound, music genre, instrument recognition, singing vocals, rhythm tempo, melody harmony, scene place, background foreground, temporal order, duration timing, overlap simultaneity, multi-source integration, phonetic pronunciation, word sound form, and other subskill.

## 3. DATASET CONSTRUCTION

### 3.1. Source Dataset

We construct a large-scale skill-annotated AQA dataset named SkillBench-AQA. The dataset is built from a large-scale AudioMCQ corpus [8], which contains multiple-choice audio question answering samples from diverse audio domains, including speech, environmental sounds, music, temporal audio events, and complex audio reasoning.

The original data contain audio clips, questions, answer options, and correct answers. However, the original labels are not sufficient for detailed skill-level analysis. Therefore, we annotate each question with a skill label and a sub-skill label.

Table 1: Skill taxonomy used in our system.

| Skill | Description |
|---|---|
| speech_content | Understanding spoken words, dialogue, statements, and semantic content. |
| speaker_identity | Recognizing speaker identity, speaker count, gender, voice identity, and same/different speaker. |
| emotion_prosody | Understanding emotion, tone, attitude, prosody, intonation, sarcasm, and speaking style. |
| temporal | Reasoning about event order, before/after relations, start/end, overlap, duration, and timing. |
| sound_event | Understanding non-speech sound events, environmental sounds, actions, and acoustic events. |
| music_style | Recognizing music genre, instruments, rhythm, melody, vocals, and musical style. |
| phonetic | Understanding pronunciation, syllables, phonemes, rhyme, and word sound form. |
| acoustic_scene | Understanding scene, background, and acoustic environment. |
| multi_source_reasoning | Combining multiple audio clues or sources. |
| other | Questions not covered by the above skills. |

### 3.2. Skill Annotation

For each multiple-choice question, the question text, candidate answers, and available metadata are used for skill annotation. The annotation process consists of three steps.

First, a large language model is used to assign one primary skill label and one sub-skill label for each sample. Second, invalid labels are filtered according to the predefined skill and sub-skill sets. Third, ambiguous samples are checked and corrected manually or assigned to the other category [9, 2].

The goal of this annotation process is not only to provide labels for analysis, but also to support skill-aware training. With these labels, we can train skill-specific adapters and construct a router that selects the most suitable skill branch for each question.

### 3.3. Dataset Statistics

The final SkillBench-AQA dataset contains 524,737 question-answer pairs and 365,670 audio clips. The dataset covers 10 main skills and 34 sub-skills. Detailed statistics are shown in Table **??**.

## 4. TRAINING

### 4.1. Baseline Model

Our system is based on MOSS-Audio-4B, a recent audio language model designed for audio understanding and instruction following [7]. We use the original instruction-tuned model as the baseline and evaluate it on the DCASE 2026 Task 5 development set.

Table 2: Skill distribution of the strongly-present subset (DeepSeek-labeled 232K samples).

| Skill | Number of samples | Ratio (%) |
|---|---|---|
| emotion_prosody | 62774 | 27.01 |
| sound_event | 45896 | 19.75 |
| speech_content | 35178 | 15.14 |
| speaker_identity | 34716 | 14.94 |
| temporal | 27114 | 11.67 |
| music_style | 11966 | 5.15 |
| multi_source_reasoning | 7406 | 3.19 |
| acoustic_scene | 7113 | 3.06 |
| phonetic | 166 | 0.07 |
| other | 60 | 0.03 |
| Total | 232389 | 100.00 |

Table 3: Full-data supervised fine-tuning configuration on AudioMCQ.

| Setting | Value |
|---|---|
| Base model | MOSS-Audio-4B-Instruct |
| Training method | Supervised Fine-Tuning (SFT) |
| Data size | 524,737 samples |
| Skill labels | 10 skills (232,389 labeled strong subset) |
| Epochs | 3 |
| Learning rate | 5e-6 |
| Micro batch size | 1 per GPU |
| GPUs | 4 |
| Gradient accumulation | 4 |
| Effective batch size | 16 |
| LoRA rank | 8 |
| LoRA alpha | 16 |
| LoRA dropout | 0.05 |
| Save interval | 500 steps |
| Log interval | 5 steps |

All questions are converted into a multiple-choice format. During inference, the model is asked to select one answer from the given options. The final answer is mapped to the corresponding option index for evaluation.

### 4.2. Full-data Supervised Fine-Tuning

We first conduct supervised fine-tuning using the full training set. The purpose of this stage is to adapt MOSS-Audio-4B to the DCASE 2026 AQA format and improve its general audio question answering ability. The training configuration is shown in Table **??**.

### 4.3. Skill-specific Adapters

The full-data SFT results show that different skills have different behaviors during training. Some skills improve, while others decrease. Therefore, we further train skill-specific adapters.

For each skill, we collect the corresponding samples from SkillBench-AQA and train an adapter for that skill. During inference, the skill router predicts the skill of the input question and selects the corresponding adapter.

Table 4: Overall development-set performance of the submitted system.

| Metric | Value |
|---|---|
| Accuracy | 56.69% |
| Correct | 911 |
| Total | 1607 |
| Valid Choice Rate | 100.00% |

### 4.4. Skill Router

To automatically select the appropriate adapter, we train a skill router. The router takes the question and answer options as input and predicts one of the ten skill labels. The predicted label is then used to select the corresponding skill adapter.

This design allows the system to use different model branches for different audio understanding skills, instead of forcing all skills to share the same fine-tuned parameters.

### 4.5. GRPO Training

We are also training a reinforcement learning stage based on Group Relative Policy Optimization (GRPO). Since the experiments are still in progress, the final GRPO results are not included in the current version. If the GRPO stage improves the development performance, it will be added as the final training stage.

## 5. EXPERIMENTAL RESULTS

### 5.1. Overall Performance

The submitted system is based on MOSS-Audio-4B-Instruct with a skill-aware routing strategy. A DeBERTa-v3 classifier is first used to predict the required skill from the question text and multiple-choice options. Two skill-specific LoRA adapters are currently deployed for `emotion_prosody` and `speaker_identity`. Questions assigned to all remaining skills are answered by the original MOSS-Audio-4B-Instruct model.

Table 4 reports the overall performance on the DCASE 2026 Task 5 development set.

The submitted system achieves an overall accuracy of 56.69% on the development set while maintaining a 100% valid choice rate.

### 5.2. Skill-level Performance

To better understand model behavior, we further evaluate performance under different skill categories. The results are shown in Table 5.

The results reveal substantial differences among skills. Speech-related questions achieve the highest accuracy among major skill categories, while temporal reasoning remains particularly challenging. Speaker identity and emotion-related questions also exhibit lower accuracy compared with speech content understanding, indicating that fine-grained audio perception remains a difficult problem for current audio language models.

Although the performance on `multi_source_reasoning` appears very high, the number of samples in this category is extremely limited and therefore does not necessarily indicate strong generalization capability.

Table 5: Skill-level performance on the development set.

| Skill | Accuracy (%) |
|---|---|
| speech_content | 74.64 |
| acoustic_scene | 62.16 |
| sound_event | 53.63 |
| music_style | 51.66 |
| emotion_prosody | 50.00 |
| phonetic | 50.37 |
| other | 50.00 |
| speaker_identity | 45.59 |
| temporal | 43.28 |
| multi_source_reasoning | 100.00 |

### 5.3. Effect of Full-data SFT

We first conducted full-data supervised fine-tuning on the complete AudioMCQ training set containing 524,737 samples. However, we observed that the overall performance improvement on the development set was limited.

To further investigate this phenomenon, we analyzed model behavior at the skill level. Our observations indicate that full-data SFT does not improve all skills uniformly. Instead, certain skills benefit from additional training while others experience performance degradation. As a result, the positive and negative effects partially cancel each other out, leading to only marginal changes in overall accuracy.

This observation motivates the use of skill-aware training strategies. Rather than forcing all audio understanding skills to share the same parameter updates, skill-specific optimization may allow the model to improve weaker skills without sacrificing performance on stronger ones.

### 5.4. Skill Router and Adapter Results

Based on the above observations, we introduce a skill-aware routing framework. A DeBERTa-v3 classifier is trained to predict skill labels from question text and answer options. The predicted skill is then used to select an appropriate adapter.

In the current system, two LoRA adapters are deployed:

- emotion_prosody adapter
- speaker_identity adapter

Questions belonging to all other skills are answered using the original MOSS-Audio-4B-Instruct model.

This design serves as an initial exploration of skill-aware training for audio question answering. Future work will extend the framework to additional skills and investigate whether larger coverage of skill-specific adapters can further improve performance.

### 5.5. Future GRPO Training

In addition to supervised fine-tuning, we are currently exploring Group Relative Policy Optimization (GRPO) for Audio Question Answering. Since these experiments are still ongoing, GRPO results are not included in the current submission.

Future work will investigate whether reinforcement learning can further improve difficult skills such as temporal reasoning, speaker identity understanding, and emotion perception.

## 6. CONCLUSION

In this technical report, we describe our submission system for DCASE 2026 Task 5: Audio Question Answering. The system is built upon MOSS-Audio-4B-Instruct and incorporates a skill-aware routing strategy based on a DeBERTa-v3 classifier.

To better understand audio language model capabilities, we construct SkillBench-AQA, a large-scale skill-annotated Audio Question Answering dataset containing 524,737 question-answer pairs, 365,670 audio clips, 10 skills, and 34 sub-skills. Based on this dataset, we perform skill-level analysis and observe substantial performance differences across audio understanding skills.

Our experiments suggest that full-data supervised fine-tuning alone may not be sufficient for balanced audio understanding. Different skills exhibit different learning behaviors, motivating the use of skill-specific adapters and routing mechanisms.

The current submission deploys dedicated adapters for emotion_prosody and speaker_identity while using the base model for all remaining skills. Future work will extend skill-aware training to additional skills and investigate reinforcement learning methods such as GRPO for further performance improvement.

## 7. REFERENCES

[1] 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.

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

[3] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[4] 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.

[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] 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.

[7] C. Yang, C. Yu, H. Chen, J. Zhu, J. Chen, K. Chen, W. Wang, Y. Wang, Y. Jiang, Y. Jiang, *et al.*, "Moss-audio technical report," *arXiv preprint arXiv:2606.01802*, 2026.

[8] 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.

[9] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.