AUDIO-GROUNDED HARD-EXAMPLE TRAINING WITH ACOUSTIC TAGGING FOR AUDIO-DEPENDENT QUESTION ANSWERING

Technical Report

Hyun Jun Kim[1,*], *Byeongchan Kim*[1,*], *Jung Chan Ryu*[2], *Yu Ra Kim*[2], *Yuri Oh*[3], *Bo Eun Choi*[3]
Changwon Lim[1,2,3,†], *Il-Youp Kwak*[1,2,3,†]

[1] Department of Smart Cities, Chung-Ang University, Seoul, Korea
[2] Department of Statistics and Data Science, Chung-Ang University, Seoul, Korea
[3] Department of Applied Statistics, Chung-Ang University, Seoul, Korea
{hyunjun0615, moch1996, jungchan0202, mormongu, yroh7736, bony1105, clim, ikwak2}@cau.ac.kr

ABSTRACT

This report describes our system for DCASE 2026 Task 5, Audio-Dependent Question Answering (ADQA), which evaluates whether models answer multiple-choice questions by grounding their decisions in the input audio rather than relying on textual priors. Our systems are built around MOSS-Audio-8B-Thinking and combine prompt calibration, audio-contrastive hard-example selection, GDPO-based LoRA adaptation, acoustic tagger-assisted inference, and permutation-based ensemble voting. We first calibrate a constrained answer-letter prompt to provide a stable interface for zero-shot inference, likelihood-based diagnostics, and training data construction. We then compare option-level scores under the original audio and a silent control signal to identify shortcut-prone examples, audio-evidence anchors, and audio-dependent error cases. These examples are used to construct a hard-example reservoir for GDPO-based LoRA adaptation. At inference time, we use a lightweight rule-routed acoustic tagger that selectively injects weak evidence from specialized analyzers, including VAD/silence features, ASR and word-level prosody, speaker diarization, speaker verification, and CLAP-based music/background matching. We submitted four systems: a training-free MOSS system with acoustic tagging, an RL-adapted MOSS system with the same tagger, a ten-permutation MOSS ensemble using majority voting, and a mixed MOSS–Qwen ensemble using four MOSS-8B and two Qwen-30B predictions under the 100B system-size limit. For the mixed ensemble, option-level log probabilities are normalized within each prediction and combined by a weighted sum. The best single MOSS-based system achieves 69.695% accuracy on the development set, while the best MOSS–Qwen ensemble achieves 70.504%.

Index Terms— Audio-dependent question answering, large audio-language models, hard-example mining, acoustic tagger

1. INTRODUCTION

Audio question answering (AQA) requires a model to infer an answer from both an audio signal and a natural-language question, making it a query-conditioned audio understanding task [1, 2]. Unlike audio captioning or closed-set audio classification, the relevant evidence is determined by the question and may involve speech content, speaker attributes, music characteristics, environmental events, temporal structure, or their interactions. This query-conditioned nature makes AQA a useful testbed for Large Audio-Language Models (LALMs), but recent studies show that audio reasoning remains difficult, particularly when the answer requires temporal evidence or when models can exploit text-only shortcuts [3, 4].

DCASE 2026 Task 5 focuses on this issue through Audio-Dependent Question Answering (ADQA), which emphasizes questions whose answers should depend on the provided audio signal [5]. The task is therefore not only about answering audio-related questions, but about suppressing text-only shortcuts and encouraging decisions that are grounded in acoustic evidence. To construct such an evaluation setting, the task uses Audio-Dependency Filtering (ADF), including silent-audio filtering, common-sense checks, soft text-shortcut filtering, and manual verification [5]. In this setting, a strong system should not merely be a capable language reasoner; it should remain sensitive to the presence, absence, and discriminative content of the audio.

We build our system on MOSS-Audio-8B-Thinking, a unified audio-language model designed for speech, environmental sound, music understanding, time-aware question answering, and audio-grounded reasoning [6]. Our main idea is to use the model's own sensitivity to audio as a training signal. We first calibrate a compact answer-letter prompt for stable multiple-choice prediction, and then compare the model's option scores under the original audio and a silent control signal. This audio-contrastive diagnostic pass is used to mine examples where the model is likely to rely on textual shortcuts, confuse acoustically plausible distractors, or fail despite useful audio evidence. We then construct a hard-example reservoir from these selected examples and adapt the model with GDPO-based LoRA training. Separately from the training pipeline, we also explore router-assisted

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No. RS-2026-25477127 & No. RS-2024-00360176).

*These authors contributed equally to this work.

†Corresponding authors: Changwon Lim and Il-Youp Kwak.

```
You are an expert in audio understanding.
IMPORTANT:
- You MUST listen to the audio carefully before answering.
- You MUST rely on the audio content to answer the question.
- Do NOT answer based only on the text of the question or
choices.
- Answer with ONLY one letter: {valid letters}.
Question: {question}
Choices:
{choice lines}
Answer:
```

Figure 1: Prompt template used as the shared answer-letter interface for direct inference, audio-contrastive log-probability diagnostics, RL rollouts, and choice-permutation scoring.

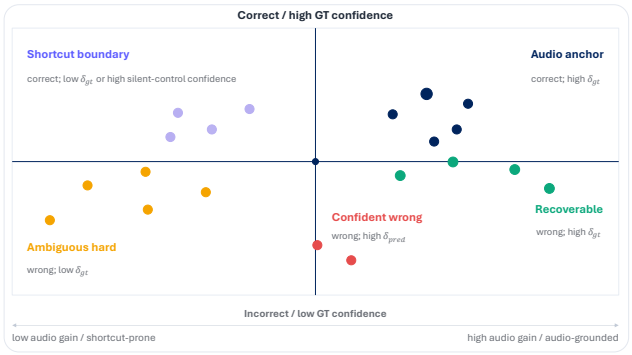

Figure 2: Conceptual boundary space for audio-contrastive hard sample mining. The horizontal axis represents audio-vs-silent gain, while the vertical axis represents correctness and confidence with respect to the ground-truth (GT) answer. Confident wrong examples are incorrect with low GT confidence, but the model-selected distractor receives high audio-conditioned support.

acoustic evidence as an inference-time module that can optionally provide weak cues from specialized audio analyzers.

Based on these components, we construct four submitted systems that progressively combine training-free MOSS inference, GDPO-adapted MOSS inference, choice-permutation voting, and mixed MOSS–Qwen log-probability fusion. The best single MOSS-based system achieves 69.695% development accuracy, and the mixed ensemble reaches 70.504%, suggesting that calibrated prompting, audio-contrastive adaptation, acoustic evidence injection, and ensemble aggregation provide complementary gains under the ADQA setting.

2. METHOD

2.1. Prompt calibration for answer-letter prediction

Before any training, we calibrated the multiple-choice prompt used by MOSS-Audio-8B-Thinking. Because large language models can be sensitive to prompt format and answer-token biases, we treat this step as a lightweight but important form of task adaptation [7, 8]. The prompt defines the model's final-answer interface and determines how answer choices are represented in the next-token distribution.

In preliminary comparisons, final answer-letter prediction was more stable than free-form answer-text generation. We therefore use the compact answer-letter interface shown in Fig. 1. This design reduces lexical variation and parsing ambiguity, fixes a common answer cue for option-level log-probability extraction, and makes choice-permutation aggregation straightforward. At inference time, for routed examples, auxiliary acoustic evidence is inserted above the question, but the final-answer interface remains unchanged: the model must select one valid option letter, which is then mapped back to the corresponding answer text for submission.

2.2. Audio-contrastive hard-example selection

Let an example be $x_i = (a_i, q_i, C_i, y_i)$, where a_i is the audio, q_i is the question, $C_i = \{c_i^A, c_i^B, c_i^C, c_i^D\}$ is the choice set, and $y_i \in \{A, B, C, D\}$ is the reference answer. We also define a silent control audio a_i^0. For each condition $r \in \{\text{audio}, \text{silent}\}$, the model assigns next-token scores $s_i^r(k) = \log P_M(k \mid a_i^r, q_i, C_i)$ for $k \in \{A, B, C, D\}$. For a target answer t, we compute the margin $m_i^r(t) = s_i^r(t) - \max_{k \neq t} s_i^r(k)$ and the audio-conditioned margin gain $g_i(t) = m_i^{\text{audio}}(t) - m_i^{\text{silent}}(t)$. In addition to these diagnostic option scores, we run real-audio generation with the calibrated answer-letter prompt and parse the final answer letter as $\hat{y}_i$. Correctness is defined by $z_i = \mathbf{1}[\hat{y}_i = y_i]$, and option confidence by $p_i^r(t) = \text{softmax}(s_i^r)(t)$.

For compact notation, let $\delta_i^{\text{gt}} = g_i(y_i)$ denote the audio gain for the reference answer and $\delta_i^{\text{pred}} = g_i(\hat{y}_i)$ denote the audio gain for the generated answer. We group examples using correctness, δ_i^{gt}, δ_i^{pred}, and silent-control confidence. Figure 2 illustrates the conceptual boundary space. Correct examples with high δ_i^{gt} become audio anchors. Correct examples with low δ_i^{gt} are treated as shortcut-boundary examples, with high $p_i^{\text{silent}}(y_i)$ marking a stronger shortcut-prone subcase. Incorrect examples are divided into recoverable examples with high δ_i^{gt}, confident wrong examples with high δ_i^{pred}, and ambiguous hard examples that do not fall into either of those two more specific failure modes. The remaining correct examples are retained as coverage examples.

The boundary thresholds are selected using quantiles of the diagnostic statistics rather than tuned on the development set. Low- and high-gain regions are defined from the empirical distributions of δ_i^{gt} and δ_i^{pred}, and confidence-based gates are defined from the silent- and audio-conditioned option probabilities.

We convert the mined groups into the RL training dataset, as described in Table 1. Rather than collecting only incorrect examples, the dataset deliberately mixes failure-focused groups with coverage-preserving and audio-anchor examples. Ambiguous hard, confident wrong, and recoverable examples expose different audio-grounded failure modes, whereas coverage examples preserve general multiple-choice behavior and audio anchors retain cases where the base model already uses audio evidence correctly. This mixture is intended to improve audio-grounded reasoning without overfitting the RL stage to narrow or pathological exam-

Table 1: Audio-contrastive group composition of AudioMCQ Strong-AC and the final RL training dataset.

Group	AudioMCQ Strong-AC	RL rows
Coverage	79,957	4,096 (32%)
Ambiguous hard	34,908	3,200 (25%)
Audio anchor	32,746	1,920 (15%)
Shortcut boundary	11,016	1,024 (8%)
Confident wrong	7,992	1,920 (15%)
Recoverable	922	640 (5%)
Total	167,541	12,800 (100%)

ples.

The final RL training dataset contains 12,800 rows selected from 167,541 AudioMCQ Strong-AC examples. We apply group-level quotas and sample within each group approximately in proportion to the square root of each source-dataset-by-question-type cell size so that large cells do not dominate the selected data. Group-specific repeat caps are also used to prevent excessive duplication. The selected dataset remains dominated by speech-related examples. However, this sampling increases the relative share of AudioCaps and sound-type questions while preserving speech-dominated coverage.

2.3. GDPO-based LoRA adaptation

We use the mined groups to construct an RL training dataset for adapting MOSS-Audio-8B-Thinking. Each training row contains the audio and the calibrated multiple-choice prompt, but no assistant gold response. The reference answer letter is retained only as a solution field for external reward computation. During training, the policy samples multiple completions for each prompt, and external reward functions evaluate whether the final answer letter is correct and whether the output follows a parseable answer format.

The reward is defined as

$$R = R_{\text{acc}} + 0.5R_{\text{fmt}}, \tag{1}$$

where R_{acc} rewards a completion whose final answer letter matches the gold label, and R_{fmt} rewards a valid final-answer format. We use a group-relative policy optimization objective with GDPO-style reward scaling across sampled completions [9, 10]. The RL stage is therefore driven by two components: audio-contrastive data selection and simple rule-based rewards for answer correctness and output validity.

2.4. Acoustic tagger-assisted inference

We use an inference-time acoustic tagger to provide question-specific acoustic evidence without modifying the MOSS-Audio parameters. The tagger consists of a question-aware router and a set of pretrained auxiliary analyzers. Given the question and answer choices, the router selects the type of acoustic cue that may be useful, but it never predicts the answer option itself. If a route is activated, the corresponding analyzer output is converted into a short textual evidence snippet and inserted into the prompt.

The acoustic router is a lightweight rule-based module. It normalizes the question and choices and applies high-precision lexical and phrase patterns for the cue types summarized in Table 2, including pause and duration cues, word-level prosody, speaker turns, same-speaker verification, music details, and background scenes. When multiple cues match, a fixed priority order reduces route conflicts, with more specialized speaker and scene routes taking precedence over broader prosody or voice activity detection (VAD) routes. The router uses only the surface form of the question and choices and does not use ground-truth labels or manual annotations of the evaluation set.

The active routes rely on standard pretrained analyzers and signal-processing tools for acoustic feature extraction, speech timing, diarization, speaker verification, and audio–text matching [11, 12, 13, 14, 15, 16]. Table 2 summarizes the analyzers used in the submitted tagger. For example, Whisper is used within the prosody route to obtain word timing, while CLAP-based routes compare the input audio with choice-specific textual descriptions for music and background-scene questions.

The analyzer outputs are heterogeneous, but we do not fuse them as internal model features. Instead, each output is normalized into a short natural-language block, such as speech and silence statistics, word-level timing and pitch cues, speaker segments, speaker similarity scores, or CLAP similarity scores. This block is inserted into the prompt as weak auxiliary analysis. The prompt explicitly states that the evidence may be imperfect and instructs MOSS-Audio to prioritize the audio itself if the auxiliary evidence conflicts with direct perception. Thus, the tagger acts as a prompt-level evidence injection mechanism rather than an ensemble vote or answer replacement.

For no-analyzer cases in the training-free MOSS branch, we use a no-auxiliary prompt selector that chooses between the calibrated base prompt and a reasoning-ban variant. The reasoning-ban prompt is used for structured semantic cases, whereas audio-perceptual and ambiguous cases keep the base prompt. This selector is not applied to the RL-adapted MOSS branch, where the original acoustic router performed better.

3. EXPERIMENTS

3.1. Experimental setup

All RL variants use the 12,800-row RL training dataset constructed by the audio-contrastive mining procedure described in Sec. 2.2. As summarized in Table 1, the selected data mixes targeted hard examples with coverage-preserving and audio-anchor examples rather than using only incorrect samples.

For RL variants with choice-layout calibration, we additionally control the target answer position during data construction. Because preliminary inference analysis showed under-production of option D, we sample the target answer position with weights A/B/C/D = 1.0/1.0/1.0/1.3 and apply a guardrail that keeps the final D ratio within 28–32%. We then reorder the answer choices so that the gold answer appears at the sampled target position. The resulting target distribution is A/B/C/D = 22.77/23.70/23.00/30.53%.

Table 2: Pretrained analyzers used by the acoustic tagger-assisted inference module.

Route / acoustic cue	**Auxiliary analyzer**	**Injected evidence**
VAD / silence / duration	librosa-based signal features [11]	Speech/silence timing and pauses.
Prosody / stress / intonation	Whisper with acoustic features [12]	Word timing and prosodic cues.
Speaker turn changes	pyannote speaker diarization [13]	Speaker segments and turn count.
Same-speaker verification	ECAPA-TDNN embeddings [14, 15]	Speaker similarity scores.
Music detail	CLAP audio–text matching [16]	Music-choice similarity scores.
Background scene	CLAP audio–text matching [16]	Background-scene similarity scores.

Table 3: Final submitted systems. Accuracy is computed over 1,607 development examples.

ID	System	Correct	Acc. (%)
S1	Pretrained MOSS + tagger + no-aux prompt router	1119	69.633
S2	RL MOSS + tagger	1120	69.695
S3	S1×5 + S2×5, hard voting	1125	70.006
S4	S1×2 + S2×2 + Qwen-30B×2, weighted log-prob fusion	1133	**70.504**

We adapt MOSS-Audio-8B-Thinking with LoRA adapters [17]. The representative run uses LoRA rank 16, LoRA alpha 32, dropout 0.0, and all linear modules as target modules. RL training uses eight generations per prompt, per-device batch size 4, gradient accumulation 2, four GPUs, 400 optimization steps, learning rate 7×10^{-7}, cosine scheduling with warmup ratio 0.01, maximum sequence length 2048, maximum completion length 512, bfloat16 precision, FlashAttention, and DeepSpeed ZeRO-2. Rollout generation uses temperature 1.0, top-k 50, and top-p 0.9.

3.2. Final submissions

We evaluate the final submitted systems on the 1,607-example development set using Top-1 accuracy. Since the calibrated models produce answer letters while the official output format requires plain answer text, post-processing maps each predicted letter to the corresponding choice text and removes residual option prefixes.

We construct the submissions in two stages. First, we build two lightweight MOSS-only systems: a training-free system based on the pretrained MOSS weights, and an RL-adapted system using the GDPO-trained LoRA adapter. Both systems use the acoustic tagger-assisted inference strategy described in Sec. 2.4, with the no-auxiliary prompt-routing policy applied only to the training-free branch.

For the non-lightweight submissions, we reuse these two MOSS branches as ensemble sources. The MOSS-only ensemble aggregates five choice-permutation views from each branch by hard voting over the predicted answer letters. The mixed ensemble further adds Qwen3-Omni-30B [18] as an inference-only branch. Qwen3-Omni-30B is not additionally trained; instead, it uses the same answer-letter interface, choice permutation, and option-level log-probability extraction as MOSS. The final mixed system combines four MOSS views and two Qwen3-Omni-30B views using weighted option-level log probabilities, resulting in an effective model-size total of 92B, which remains below the 100B system-size limit.

For each valid view v, we first map the option log-probability vector to the original choice order. Let V_M and V_Q denote the valid MOSS and Qwen views, respectively. We compute branch-level scores by averaging log probabilities within each branch:

$$s_k^M = \frac{1}{|V_M|} \sum_{v \in V_M} l_{v,k}, \qquad s_k^Q = \frac{1}{|V_Q|} \sum_{v \in V_Q} l_{v,k}.$$

The final answer is selected by

$$\hat{y} = \arg\max_k \left(s_k^M + \lambda s_k^Q \right), \qquad \lambda = 0.46.$$

It should be noted that S1 and S2 are different single-system configurations rather than a controlled ablation of RL alone. S1 uses the pretrained MOSS model with acoustic tagging and a no-auxiliary prompt-routing policy, whereas S2 uses the GDPO-adapted MOSS model with the acoustic tagger. Thus, the S1—S2 comparison should be interpreted as a system-level comparison between the training-free and RL-adapted branches, not as the isolated contribution of RL. The contribution of the RL branch is further reflected in the ensemble submissions, where S2 is used as one of the MOSS sources.

The best single MOSS-based submission is S2, which uses the RL-adapted model with acoustic tagger-assisted inference. The best overall submission is S4, where weighted log-probability fusion improves the development accuracy to 70.504%.

4. CONCLUSION

We present a compact ADQA system centered on MOSS-Audio-8B-Thinking. The system combines calibrated answer-letter prompting, audio-contrastive hard-example selection, GDPO-based LoRA adaptation, acoustic tagger-assisted inference, and permutation-based aggregation. Among the submitted systems, the best single MOSS-based configuration reaches 69.695% development accuracy, and the mixed MOSS–Qwen weighted log-probability ensemble reaches 70.504%.

5. REFERENCES

[1] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," in *Proceedings of the 30th European Signal Processing Conference* (*EUSIPCO*), 2022.

[2] P. Sudarsanam and T. Virtanen, "Attention-based methods for audio question answering," 2023.

[3] A. K. Sridhar, Y. Guo, and E. Visser, "Enhancing temporal understanding in audio question answering for large audio language models," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Industry Track*, 2025.

[4] 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 *International Conference on Learning Representations* (*ICLR*), 2026.

[5] "DCASE 2026 challenge task 5: Audio-dependent question answering," https://dcase.community/challenge2026/task-audio-dependent-question-answering, 2026, accessed: 2026-06-04.

[6] C. Yang, C. Yu, H. Chen, J. Zhu, J. Chen, K. Chen, W. Wang, Y. Wang, Y. Jiang, Y. Jiang, Z. Lin, Z. Chen, Z. Fei, C. Liu, J. Zhan, K. Yu, K. Huang, M. Chen, Q. Cheng, R. Li, S. Li, S. Wang, Y. Gao, Y. Zhang, and X. Qiu, "Moss-audio technical report," 2026.

[7] T. Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, "Calibrate before use: Improving few-shot performance of language models," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 12 697–12 706.

[8] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr, "How i learned to start worrying about prompt formatting," in *The Twelfth International Conference on Learning Representations*, 2024.

[9] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "DeepSeekMath: Pushing the limits of mathematical reasoning in open language models," 2024.

[10] S.-Y. Liu, X. Dong, X. Lu, S. Diao, P. Belcak, M. Liu, M.-H. Chen, H. Yin, Y.-C. F. Wang, K.-T. Cheng, *et al.*, "Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization," *arXiv preprint arXiv:2601.05242,* 2026.

[11] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–24.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 28 492–28 518.

[13] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proceedings of INTERSPEECH 2023*, 2023, pp. 1983–1987.

[14] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proceedings of INTERSPEECH 2020*, 2020, pp. 3830–3834.

[15] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624,* 2021.

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

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

[18] 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.