TASK-LEAF ROUTED MIMO-AUDIO FOR DCASE 2026 TASK 5

Technical Report

Jongha Kim, Leehyeon Song*, Hyung-Min Park*†

Intelligent Information Processing Lab
Sogang University
Seoul, Korea
{hytric, dlgus53, hpark}@sogang.ac.kr

ABSTRACT

This report describes our submission system for DCASE 2026 Task 5, Audio-Dependent Question Answering (ADQA). ADQA is a multiple-choice audio question answering task in which a system receives an audio clip, a question, and answer choices, and returns the plain-text answer choice using acoustic evidence. Our system uses MiMo-Audio-7B-Instruct as a shared backbone and applies a task-leaf routed design. A text-only router assigns each sample to a fine-grained task leaf, and a fixed leaf-level decision rule selects a base two-pass Chain-of-Thought (CoT) path, a Supervised Fine-Tuning (SFT) LoRA adapter, or a Group Relative Policy Optimization (GRPO) evidence route. This design is motivated by broad adaptation results: on the 1607-sample development set, broad SFT and broad GRPO underperform the base two-pass path overall, while improving selected leaves. The submitted routed system obtains 1180/1607 correct answers, or 73.430% submitted development accuracy. This is a submitted development metadata score, not a hidden evaluation score or a single always-on model result.

***Index Terms*—** ADQA, LALM, task routing, LoRA, GRPO

1. INTRODUCTION

DCASE 2026 Task 5 evaluates Audio-Dependent Question Answering (ADQA), a setting in which a system receives an audio clip alongside a natural-language question and a fixed set of answer choices, and is required to return the exact plain text of the correct answer [1]. Building on recent audio question answering benchmarks and the DCASE 2025 Task 5 setting, this task places particular emphasis on audio dependency by requiring that the intended solution not be recoverable from question text, commonsense priors, option wording, option position, or output format alone [2, 3].

Although Large Audio-Language Models (LALMs) are capable of jointly processing speech, environmental sound, and music with natural-language instructions, multiple-choice ADQA exposes several failure modes that remain less salient in open-ended audio captioning. A model may exploit plausible answer priors, reproduce an option letter rather than the corresponding answer text, or generate a reasoning span that violates the required submission format—risks that are consistent with prior studies on text bias in LALMs and option-position sensitivity in multiple-choice evaluation [4, 5, 6]. These failure modes are especially pronounced for question types involving temporal boundaries, duration estimation, emotion and tone recognition, phonetic reasoning, source identification, and speaker profiling, all of which demand fine-grained acoustic cues that cannot be inferred from surface-level linguistic patterns.

Our system is motivated by the observation that a single global adaptation strategy is insufficient to serve all task types uniformly: while broad LoRA SFT and broad GRPO improve performance on certain music-related leaves, they degrade accuracy on speech and temporal groups on the development set. To address this, we adopt a task-leaf router combined with a leaf-level decision rule, where the router is responsible solely for assigning the task leaf from the question and answer choices, and the decision rule independently selects the validated backend for each leaf. This decomposition separates task identification from model selection, thereby preventing a universally applied adapter from overriding the base inference path in cases where adaptation yields no benefit.

2. METHOD

2.1. System overview

The submitted system takes an audio clip, a question, and a set of answer choices as input, and produces as output the plain text of exactly one answer choice. At the first stage, a text-only task-leaf router assigns each sample to a fine-grained task leaf solely from the question and choices, without accessing the audio signal at inference time; the assigned leaf is then passed to a fixed decision rule that selects the appropriate backend, with audio introduced only at this subsequent stage.

Figure 1 summarizes the overall pipeline. The default backend performs MiMo-Audio two-pass chain-of-thought (CoT) inference, while validated leaves may instead activate a task-specific SFT LoRA adapter or a GRPO evidence adapter whose generated evidence is forwarded to a base selector. Regardless of which backend is activated, all paths share a common output contract: an answer normalization layer maps raw model output to exactly one of the provided choices, ensuring that option letters, explanations, or evidence spans are excluded from the submission CSV.

2.2. Data curation

A fundamental challenge in this setting is the distributional mismatch between the provided training pool and the development set. The official training resource is derived from AudioMCQ, a large audio multiple-choice question dataset with audio-contribution-aware filtering [3], whereas the development set draws on recent

*Co-authors. †Corresponding author.

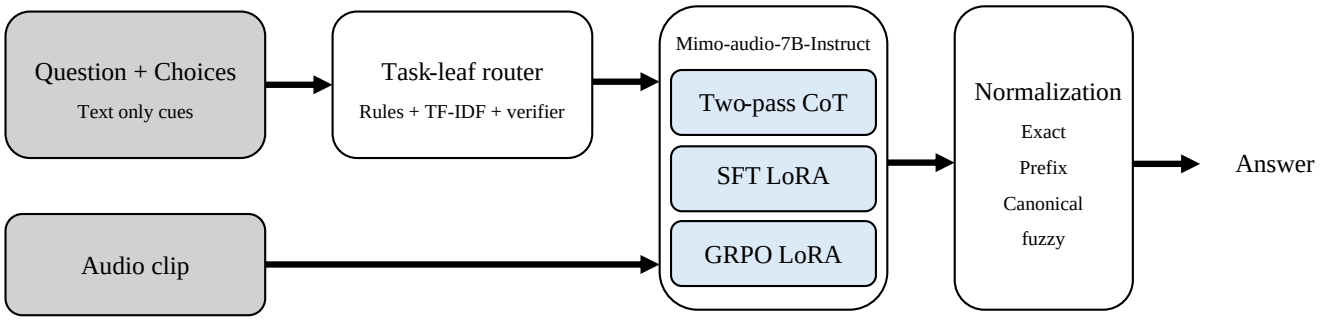

Figure 1: Overview of the task-leaf routed MiMo-Audio system. The router uses only text-side cues to predict a task leaf, while the audio signal is consumed only by the selected backend. All backend outputs are normalized to the required plain answer text.

audio understanding and reasoning benchmarks including MMAU, MMAR, and MMSU [7, 8, 9]. Naïvely fine-tuning on the full training pool risks overfitting to question types already well-handled by the base model, while leaving underrepresented categories—such as phonetics, speech boundary detection, and audio quality assessment—without adequate supervision. To address this mismatch, we adopt a targeted data construction strategy: rather than filtering or reweighting the existing corpus, we directly construct category-specific training instances from source datasets, guided by the leaf taxonomy defined on the development set, thereby enabling explicit control over the category-level composition of the training data.

Central to this strategy is the organization of the development set into a two-level hierarchy of task families and task leaves, where each sample is assigned to exactly one leaf. Leaves are defined according to three principles. First, as the benchmark is partly template-generated, recurring question templates naturally cluster into tight leaves. For instance, tasks such as stop-release detection, phoneme matching, and conclusion inference repeatedly appear with highly similar question structures and require the same underlying reasoning process. We therefore organize these recurring templates into dedicated leaves based on the reasoning operation required to answer the question, rather than the acoustic content itself. Second, open-ended or free-form questions that follow no distinctive template are instead split by the acoustic domain of the content—speech, music, environmental sound, or emotion—such that free-form description questions are separated into leaves such as `musical_characteristic` or `source_characteristic` by domain rather than by wording. Third, leaves are refined post-hoc whenever an internal sub-cluster exhibits systematically different model behavior and a dedicated treatment improves accuracy; for example, `rhythm_or_tempo` was carved out of musical characteristic because rhythm and timing questions benefited from a different inference strategy than melodic-description questions, with such splits being data-driven rather than defined a priori.

Leaf-specific training instances are generated from existing dataset labels rather than new audio annotation, with source labels from MELD, AudioCaps, TACOS, and SpeechCraft converted into multiple-choice answer targets [10, 11, 12, 13]. Question text and plausible distractors are generated using Qwen3-235B [14], while deterministic checks enforce consistency across the manifest: each instance must employ four answer choices, preserve exact answer-choice matching, avoid answer-position imbalance, and pass audio-dependency checks including silent-audio and text-only solvability controls. Reference datasets used for data construction include AVQA, ClothoAQA, CompA-Order, and TACOS [15, 16, 17, 12].

Table 1: Development-set task-family distribution used to define the 68 task leaves. Share is the percentage of development rows in each family.

task family	leaves	share(%)
speech semantic	14	23.3
sound event	7	13.6
music content	9	13.5
speaker profile	7	12.8
speech emotion	6	9.2
speech recognition	8	6.7
speech delivery style	6	6.6
scene context	2	5.0
phonetics/phonology	5	4.6
speech boundary	3	4.4
audio quality	1	0.2
total	68	100%

2.3. Routing model

The router is text-only: it is conditioned solely on the question and its answer choices and never accesses the audio signal. Answer choices are sorted case-insensitively before featurization, so samples differing only in option order receive the same route. Routing then proceeds in two steps. First, a compact set of high-precision deterministic rules resolves questions whose phrasing maps unambiguously to a single leaf. The remaining samples are routed by a two-stage TF-IDF classifier: a parent classifier predicts the top-level task family, and a separate 68-way leaf classifier predicts the fine-grained leaf, which regularizes the high-cardinality leaf decision.

Let x be the text input, $\mathcal{L}$ the set of leaves, and $\pi(l)$ the parent family of leaf l. With $P_{\text{leaf}}(l \mid x)$ and $P_{\text{par}}(\pi(l) \mid x)$ the leaf and parent posteriors, we score each leaf as

$$s(l) \;=\; P_{\text{leaf}}(l \mid x)\,\big[\,P_{\text{par}}(\pi(l) \mid x)\,\big]^{\gamma}, \qquad \gamma = 0.5, \quad (1)$$

and assign $\hat{l} = \arg\max_{l \in \mathcal{L}} s(l)$. The exponent $\gamma = 0.5$ treats the parent as a soft prior rather than a hard gate: it down-weights leaves from unlikely families while still letting a confident leaf classifier override an uncertain parent prediction.

The assigned leaf selects a backend through a leaf-to-backend mapping that is fixed before evaluation and uses neither the audio signal, the sample identity, nor any held-out label. Low-confidence samples and leaves with insufficient validation evidence fall back to the base two-pass path.

Base model. The base model is MiMo-Audio-7B-Instruct, a 7B autoregressive audio-language model that consumes audio through the MiMo-Audio tokenizer [18]. We use it as the shared pretrained backbone for all submitted systems. Without any adapter, it can run a two-pass CoT inference procedure [19]. In the first pass, the model generates short audio-grounded evidence from the audio, question, and choices. In the second pass, it uses the original question, choices, and generated evidence to select the final answer text. The evidence is an inference-time intermediate context only; it is not exposed in the final output and is not used as a supervised training target.

This base two-pass path reaches 947/1607 correct answers, or 58.93%, on the 1607-sample development set. It is therefore the main fallback backend for leaves where a specialist route is not validated.

Per-leaf strategy selection. Per-leaf strategy selection uses three backend types. The first is the base two-pass CoT path, used by default. The second is SFT LoRA, used when a leaf has a clear capability deficit and leaf-specific labeled data are available [20]. The third is a GRPO evidence route, used when the model captures relevant evidence but final answer selection is unstable [21].

This rule is motivated by development-set evidence. Broad SFT and broad GRPO obtain 55.26% and 55.44%, respectively, below the 58.93% base two-pass score. Their effect is not uniform: both broad adapters improve the music group but regress speech and temporal groups. Specialist adapters are therefore activated only when a leaf-level validation gain is observed. At test time, no gold label or row identifier is used; the system uses only the predicted leaf and the pre-fixed leaf-to-backend mapping.

3. EXPERIMENTS AND RESULTS

3.1. Training and inference details

Training is formulated as multiple-choice answer-text generation. The input is audio, question, and answer choices, and the target is the plain text of the gold answer choice. SFT adapters are trained with cross-entropy on the answer text. Reasoning spans and evidence strings are excluded from the target. GRPO is also evaluated through normalized answer text, but in the submitted routed system it is described as an evidence route rather than as a universal final answer head.

Table 2 reports the common hyperparameter setting that is kept in the paper. Development included additional adapter variants, but the report intentionally presents one reproducible setting for the main MiMo LoRA SFT family. GRPO uses the same LoRA adapter dimensions as SFT; optimizer, rollout, and reward-schedule details are left out of this draft.

Post-processing is a deterministic answer-normalization stage required by the Task 5 output format. Raw model outputs may include an option letter, a partial option string, or short reasoning text. The submitted systems apply a rule-based cascade in the following order: exact option-text match, letter-prefix parsing, canonical string normalization, substring containment, and fuzzy similarity. This step does not use hidden labels. It only maps the model output to one of the visible answer choices.

Table 2: Reported training configuration for the main MiMo LoRA SFT family.

item	value
backbone	MiMo-Audio-7B-Instruct
adaptation	LoRA [20]
rank / alpha / dropout	$r = 8$, $\alpha = 32$, dropout $= 0.05$
GRPO LoRA setting	same adapter setting as SFT; detailed schedule omitted
target modules	decoder self-attention and MLP projections
objective	answer-text cross-entropy
epochs	2
learning rate	5×10^{-5}
batch / accumulation	7 / 2
random seed	45

3.2. Development setup and results

We compare one external reference system, Qwen3-Omni exact-option inference [22], with MiMo-Audio variants. MiMo variants include the provided/generated CSV baseline, base two-pass CoT, broad LoRA SFT, broad GRPO, and the deployed task-specific routing system. Results are reported on the 1607-sample development set. The category columns are aggregated from row-level split keys: speech includes speech-related, speaker-profile, and phonetics leaves; sound includes sound-event, scene/context, and audio-quality leaves; music includes music-content leaves; temporal includes speech-boundary and temporal-relation leaves. Three miscellaneous rows are included only in the all column.

Table 3 shows the central motivation for routing. Broad SFT and broad GRPO are worse than the base two-pass path overall. However, their leaf-level behavior differs by category: music improves from 48.61% with the base two-pass path to 59.72%, while speech and temporal groups regress. This pattern is consistent with inter-leaf training interference and motivates a routed design rather than a single global adapter.

Table 3: Component comparison on the 1607-sample development set. The routed row corresponds to 1180/1607 correct answers. Values are accuracy in percent.

system	speech	sound	music	temp.	all
Qwen3-Omni	67.59	66.56	61.11	59.15	66.09
MiMo two-pass	62.27	60.26	48.61	36.62	58.93
MiMo base	57.24	60.26	49.54	19.72	55.07
↪ Broad SFT	56.16	54.97	59.72	29.58	55.26
↪ Broad GRPO	56.06	55.30	59.72	33.80	55.44
↪ **Routed (ours)**	**74.78**	**69.54**	**73.15**	**71.83**	**73.43**

Table 4 shows that SFT is most useful in leaves with a capability deficit, such as chord/key/harmony and sound counting. GRPO is applied more narrowly, where evidence is available but final answer choice is unstable. In the current route map, the best route distribution over 69 split-subtask routes is 65 SFT LoRA routes, three GRPO evidence routes, and one base two-pass route. This distribution is a development-set engineering choice and should not be interpreted as an oracle for hidden evaluation.

Table 4: Representative routing effects per specific task. Each row compares the base two-pass path with the best routed backend for a validated specific task (leaf). Values are accuracy in percent.

specific task	method	base	routed	Δ
musical emotion	SFT	68.18	95.45	+27.27
chord/harmony	SFT	21.43	82.14	+60.71
speech duration	SFT	45.71	57.14	+11.43
sound counting	SFT	25.00	68.75	+43.75
pause meaning	GRPO	42.31	61.54	+19.23

The main result is that task-specific routing outperforms broad adaptation on the development set, but the interpretation is limited. The submitted routed score is derived from a fixed leaf-to-backend mapping selected from development-set evidence. If the hidden evaluation distribution changes, the same mapping may behave differently. For this reason, the result is reported as a submitted development metadata score and not as a claim about hidden evaluation performance.

The second limitation concerns GRPO. In this draft, GRPO is not presented as a global improvement method. Its useful role is a restricted evidence route for a small number of leaves. This differs from SFT LoRA, which is applied to many capability-deficit leaves after validation. The system therefore combines specialists conservatively and keeps the base two-pass path as the fallback whenever a specialist gain is unclear.

4. CONCLUSION

We described a MiMo-Audio based task-leaf routed system for DCASE 2026 Task 5. The system uses a text-only router to assign each sample to a fine-grained task leaf and then applies a fixed decision rule to select the base two-pass CoT path, an SFT LoRA adapter, or a GRPO evidence route. The design is motivated by the failure of broad SFT and broad GRPO to improve the full development set uniformly. The submitted routed system reaches 73.430% submitted development accuracy on the 1607-sample development set. The result should be read as a routed-system development score, not as a hidden evaluation score or a single-model result.

5. ACKNOWLEDGMENT

We thank the DCASE Challenge organizers for releasing the Task 5 template, data, and evaluation protocol.

6. REFERENCES

[1] DCASE Challenge, "DCASE 2026 Challenge Task 5: Audio-Dependent Question Answering," https://dcase.community/challenge2026/task-audio-dependent-question-answering, 2026, accessed: 2026-06-23.

[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] 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," in *International Conference on Learning Representations*, 2026.

[4] C. Wang, G. Deng, X. Yang, H. Qiu, and T. Zhang, "When audio and text disagree: Revealing text bias in large audio-language models," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 4878–4888.

[5] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang, "Large language models are not robust multiple choice selectors," in *International Conference on Learning Representations*, 2024.

[6] F. López, S. Kesiraju, and J. Luque, "Robustness assessment of large audio language models in multiple-choice evaluation," *arXiv preprint arXiv:2510.04584*, 2025.

[7] 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*, 2025.

[8] 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," *arXiv preprint arXiv:2505.13032*, 2025.

[9] 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*, 2026.

[10] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.

[11] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 119–132.

[12] P. Primus, F. Schmid, and G. Widmer, "TACOS: Temporally-aligned audio captions for language-audio pretraining," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2025, pp. 1–5.

[13] Z. Jin, J. Jia, Q. Zhou, H. Meng, *et al.*, "SpeechCraft: A fine-grained expressive speech dataset with natural language description," *arXiv preprint arXiv:2408.13608*, 2024.

[14] 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.

[15] P. Yang, X. Wang, X. Duan, H. Chen, R. Hou, C. Jin, and W. Zhu, "AVQA: A dataset for audio-visual question answering on videos," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3480–3491.

[16] 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*, 2022.

[17] S. Ghosh, A. Seth, S. Kumar, U. Tyagi, C. K. R. Evuru, S. Ramaneswaran, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "CompA: Addressing the gap in compositional reasoning in audio-language models," in *International Conference on Learning Representations*, 2024.

[18] Xiaomi LLM-Core Team, "MiMo-Audio: Audio language models are few-shot learners," *arXiv preprint arXiv:2512.23808*, 2025.

[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, 2022.

[20] 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*, 2022.

[21] 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," *arXiv preprint arXiv:2402.03300*, 2024.

[22] 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.

7. APPENDIX

The appendix records the prompt and routing artifacts needed to reproduce the submitted system behavior. These tables are implementation templates only and do not introduce additional evaluation claims.

```
You are answering a multiple-choice audio
question.  You will receive the raw audio, the
question, and four options.

## Your Task
Choose the best option for the given question
using the audio.

## Question
{QUESTION}

## Choices
A. {CHOICE_A}
B. {CHOICE_B}
C. {CHOICE_C}
D. {CHOICE_D}

## Critical Instruction
1.  Output only the exact text of one of the
four choices.
2.  Do not include the option letter (A/B/C/D)
or any explanation.
3.  Do not invent a new option.

## Output Format
Answer:  {exact choice text}

Now, answer the question.
```

Table 5: Direct inference prompt template.

```
You are answering a multiple-choice audio
question in two passes.  In Pass 1 you only
produce short audio-grounded evidence.

## Your Task
Listen to the audio and write short evidence
that is relevant to the question.  Do not
choose an option yet.  Do not output any
letter.

## Question
{QUESTION}

## Choices
A. {CHOICE_A}
B. {CHOICE_B}
C. {CHOICE_C}
D. {CHOICE_D}

## Output Format
Evidence:  {one to two sentences grounded in
the audio}

Now, output the evidence.
```

Table 6: Two-pass inference prompt template (Pass 1: evidence generation).

```
You are answering a multiple-choice audio
question in two passes.  In Pass 2 you select
the final answer using the audio, the question,
the choices, and the Pass 1 evidence.

## Your Task
Use the audio, the question, the choices, and
the Pass 1 evidence to select the best answer.
Output only the exact answer choice text, not
the option letter.

## Question
{QUESTION}

## Choices
A. {CHOICE_A}
B. {CHOICE_B}
C. {CHOICE_C}
D. {CHOICE_D}

## Audio-grounded evidence
{EVIDENCE}

## Critical Instruction
1.  Output only the exact text of one of the
four choices.
2.  Do not include the option letter (A/B/C/D),
evidence, or any explanation.
3.  The answer must be consistent with the
audio and the Pass 1 evidence.

## Output Format
Answer:  {exact choice text}

Now, output only the final answer.
```

Table 7: Two-pass inference prompt template (Pass 2: answer selection).