DCASE 2026 AUDIO-DEPENDENT QUESTION ANSWERING TASK

Technical Report

Aniket Tathe[1,2*]

[1] University of Illinois Urbana-Champaign, Urbana, IL, USA
[2] Carnegie Mellon University, WavLab, Pittsburgh, PA, USA

ABSTRACT

We describe the UIUC/CMU WavLab submission to DCASE 2026 Task 5, Audio-Dependent Question Answering (ADQA), an entry to the sub-10B lightweight champion track. The official training set pairs each multiple-choice question with a chain-of-thought (CoT) reasoning trace generated by Gemini 3.1 Pro, but we find these traces are systematically *contaminated* in two ways that train exactly the textual-hallucination behaviour the benchmark is built to penalize: 68.5% do not engage every answer option with discriminating evidence, and 10.1% reason from a written "transcript" that does not exist at inference time. We therefore repair the data rather than the model, regenerating the 14,129 flagged traces with a blind, audio-only Gemma-4-E4B-it→Qwen3-Omni teacher cascade under rejection sampling: a trace is kept only when the teacher independently selects the gold option from the audio alone. We supervised-fine-tune Qwen2.5-Omni-7B (8.93B parameters) on the repaired data and further optimize the best variant with GRPO; a controlled ablation shows that a reference-guided Prometheus reasoning reward does not improve accuracy, so our strongest model uses answer-accuracy and format rewards alone. The setting is deliberately low-resource: we use only the 19,480 officially provided items, a small fraction of the 570k+ samples behind comparable systems, and add no external data or models. On the official 1,607-item development set our best system reaches 58.4% top-1 accuracy, $+5.4$ points over the un-tuned base model.

***Index Terms*—** audio question answering, chain-of-thought distillation, data decontamination, reinforcement learning

1. INTRODUCTION

Large audio-language models (LALMs) increasingly answer free-form questions about speech, sound, and music, yet a recurring concern is that much of their measured accuracy comes from language priors rather than from genuine listening. DCASE 2026 Task 5, Audio-Dependent Question Answering (ADQA), is designed to isolate listening ability: every evaluation question is passed through an audio-dependency filtering (ADF) pipeline (silent-audio screening, large-language-model common-sense checks, perplexity-based soft filtering, and manual verification), so that the correct option cannot be recovered from the question and choices alone [1]. Systems are ranked by top-1 accuracy on the 3,000-question ADQA-Bench evaluation set, on which a random guess scores 25.5%.

The official training corpus, AudioMCQ-StrongAC-GeminiCoT, is drawn from the 570k+ sample AudioMCQ dataset through its StrongAC (strong audio-contribution) split, which selects items whose answer is highly dependent on audio cues. It provides 19,480 multiple-choice items, each annotated with a chain-of-thought (CoT) trace generated by Gemini 3.1 Pro and intended as distillation supervision. A deterministic audit of these traces (Section 2) reveals two pervasive defects that teach precisely the behaviour ADQA penalizes. **Q1:** in 68.5% of items the CoT fails to quote at least one of the three distractor options verbatim (in the large majority, none of them), providing little supervision for the option-discrimination skill the benchmark stresses. **Q2:** in 10.1% of items (97% of them speech) the CoT reasons from a written "transcript" or "provided text", an artifact unavailable at inference time, which trains the model to emit confident transcriptions it never heard and to cite documents it was never given.

Our guiding principle is to *repair the data, not the model*. We regenerate the 14,129 contaminated traces with an audio-only teacher cascade run under *blind rejection sampling*: the teacher is shown only the audio, question, and options (never the gold answer or the original CoT), and a regenerated trace is kept only when the teacher's independently chosen option matches the gold label, so that a correct answer is itself evidence the teacher actually listened. We then fine-tune Qwen2.5-Omni-7B on three controlled data variants that differ only in which traces are repaired, and further optimize the strongest variant with Group Relative Policy Optimization (GRPO) [2] using a reference-guided reasoning-quality reward. We reuse the training scripts of the AudioMCQ recipe [3], from which the official corpus derives, but operate at a small fraction of its scale: where AudioMCQ draws on more than 570k samples, we use only the 19,480 officially provided items and add no external data or models. This low-resource, rules-clean setting, together with a sub-10B model, places our entry in the lightweight champion track. This report contributes: (1) a reproducible contamination audit of the official CoT labels; (2) a blind rejection-sampling regeneration pipeline that produces grounded, option-discriminating traces; and (3) a controlled SFT-and-GRPO study that reports transparently both where the repaired supervision helps and where the variants are statistically indistinguishable on the development set.

2. TRAINING-DATA DECONTAMINATION

We audit all 19,480 Gemini CoT traces with a deterministic classifier, so every statistic below is reproducible.

2.1. Two contamination patterns

Q1: no engagement with the options. A trace is flagged Q1 when fewer than all three distractor options appear verbatim in it, that is, at least one distractor is never mentioned (case-insensitive

*Corresponding author: atathe2@illinois.edu

Table 1: Verified contamination statistics per question type. The Q1 and Q2 columns overlap; the union counts each flagged row once.

type	n	clean	Q1	Q2	flagged (∪)
speech	10,505	2,408	7,344	1,919	8,097
sound	6,085	1,898	4,162	39	4,187
music	1,724	583	1,138	12	1,141
temporal	1,166	462	701	7	704
total	19,480	5,351	13,345	1,977	14,129

substring). Such a trace does not contrast the full set of alternatives, so it provides little supervision for the option-discrimination skill the benchmark stresses. Because ADQA's distractors survived perplexity-based soft filtering, they are text-plausible by construction, making this discrimination the crux of the task. Engagement is in practice bimodal: a trace that mentions any distractor almost always mentions all three, so partial engagement (one or two of the three) is only 5.6% of items, and 12,249 of the 13,345 flagged traces mention no distractor at all. Q1 is pervasive across *every* category (7,344 speech, 4,162 sound, 1,138 music, and 701 temporal traces), for a total of 13,345 rows (68.5%).

Q2: reference to a written document. A trace is flagged Q2 when it makes a determiner-led reference to a "transcript", "transcription", or "provided text" (exact regular expression). These traces reason from an artifact that does not exist at inference time, training the model to emit confident transcripts as if heard and to cite a document it was never given, precisely the textual-hallucination behaviour ADF penalizes. Unlike Q1, Q2 is almost exclusively a speech phenomenon: 1,919 of the 1,977 flagged rows (10.1% overall) are speech.

Conservative criteria. Each candidate pattern was decided by printing real matched contexts and rejecting it if even one match admitted a listening-compatible reading. Timestamps are the clearest case: a phrase such as "the receiver sound at 00:06" could be a genuinely heard event time or copied from a transcript, and on temporal questions a listening model may legitimately report timing. Because surface form cannot separate the two, we leave *all* timestamps unflagged, accepting that a few contaminated temporal traces slip through rather than risk discarding grounded ones. We likewise do not flag verbatim quoted speech without a document reference (surface form cannot prove "read" vs. "heard"), or ambiguous words such as "the text", "passage", or "the prompt". A starker pattern, traces that openly admit no audio was available, is routed to a 217-row "lack-of-audio" manual-review sidecar rather than auto-flagged, because one audited case plausibly referred to a missing sound *within* a heard clip.

2.2. Illustrative traces

The excerpts below are copied verbatim from the training file. Q2 contamination concentrates in the speech subset:

- *Explicit transcript* (speech, SpeechCraft): "The transcript clearly reads, '00:00 And there is only one of me, said the snapping turtle to himself.' " It names the document and copies it verbatim, including a caption marker '00:00' the speaker never uttered.
- *Lack-of-audio sidecar* (speech, SpeechCraft): "Given my lack of audio input, I'm left to fall back on common patterns and typical datasets . . . I choose 'Male'. . . " The labeler admits no audio and answers from priors.

2.3. Blind rejection-sampling regeneration

We regenerate only the 14,129 flagged traces; the 5,351 clean rows keep their original CoT. The teacher is *blind*: it receives only the audio, question, and options (never the gold answer or the original trace) and must choose an option itself. We keep a regenerated trace *only when the teacher's independently chosen option equals the gold label* (rejection sampling), so a correct answer is itself evidence the teacher listened. This is a deliberate departure from a gold-conditioned design: a pilot showed gold-conditioned teachers confabulate, describing the same clip as having "trumpets" when shown the answer but "violins" when blind. The accept filter additionally requires that every distractor be addressed by name with a perceptual verdict, that no banned document vocabulary appear, that verbatim quotes span at most eight words, that timestamps appear only for temporal questions, and that the trace end with the exact option text wrapped in an `<answer>` tag for compatibility with the answer-accuracy reward used later.

Design intent: deliberate per-option discrimination. The accept filter is not only a cleanliness gate; it encodes the central hypothesis behind the repair. A contaminated trace asserts a winner without weighing the alternatives, so the student never learns to *rule options out* and may hallucinate a justification for whichever choice it favours. We therefore require every regenerated trace to follow an explicit elimination format: it opens with first-person perceptual evidence ("I hear . . . "), then issues a separate verdict on *each* option by its exact wording, marking a distractor `Rejected` with the perceptual reason it fails or `Uncertain` when the audio is inconclusive, before the final answer tag. Our intent was to test whether this forced per-option accounting teaches the model to assess every option rather than pattern-match a winner, yielding more structured reasoning, less confident hallucination, and direct supervision for the option-discrimination skill ADQA isolates. Because the gate keeps only traces whose independently chosen option is correct, this structure is learned exclusively from verified-grounded examples.

To control cost we run a teacher *cascade*. Gemma-4-E4B-it [4] (greedy) attempts every row and is accepted on 8,948 rows (63.3%); its 5,180 rejects escalate to the larger Qwen3-Omni-30B-A3B-Thinking [5], which together with sampling-retry fallbacks on the remainder recovers a further 3,935. Gemma is inexpensive but weak on fine perceptual distinctions such as speaker gender, instrument timbre, and sound identity, which is exactly where the larger model rescues it. The cascade accepts **12,883 of 14,129 regenerations (91.2%)**.

The remaining 1,246 rows, which no teacher answered correctly while blind even after the fallback retries, have no certifiable repair, since answer-conditioned regeneration of them was found to fabricate evidence, so we drop them entirely. This leaves a decontaminated pool of **18,234 items** that carry a verified CoT (5,351 clean plus 12,883 regenerated), every one drawn solely from the officially provided data. Dropping the same rows everywhere also removes any data-quantity confound, so the variants below differ only in CoT *content*.

2.4. From contaminated to grounded

Two repairs (training-row id in parentheses, excerpts verbatim) show the transformation; each regenerated trace reaches the same option but replaces guessing or transcript copying with first-person perception and a `Rejected`/`Uncertain` verdict per option.

Q1, speaker id (`librittsr_6482_289558`). Before: "I'm leaning towards 'An older man' [...] a common profile for this type of narration." After: "I hear a mid-to-low pitch suggesting an adult male. A teenage boy: Rejected (too mature); A young woman: Rejected (clearly male); An older man: Uncertain (cannot tell older from middle-aged); An elderly woman: Rejected (male)."
Q2, spoken content (`librittsr_7517_100437`). Before: "Based on the provided text [...] 'Samuel Butler' immediately stands out." After: "I hear the speaker state he found Samuel Butler. Charles Dickens / Ernest Hemingway / George Eliot: Rejected (not mentioned); Samuel Butler: stated explicitly."

3. SYSTEMS

3.1. Model and data partition

All systems use Qwen2.5-Omni-7B [6] in a thinker-only configuration; the speech-generation talker is disabled because the task requires only text. The inference model has 8.93B parameters, within both the 30B per-component cap and the 10B threshold for the Lite-ADQA lightweight track. We split the data 70/30 by audio identity into disjoint SFT and GRPO partitions, stratified by question type and source dataset; because each audio clip contributes a single row, no clip is shared across the two. The assembled SFT training set has 12,569 items and the GRPO set 5,665 prompts, all from the provided corpus.

3.2. Supervised fine-tuning: three controlled variants

We run full-parameter SFT (no LoRA) on three variants that share the identical 12,569-row partition and differ only in which CoTs are repaired:

- **A, baseline:** every row keeps its original Gemini CoT (0 repaired). Retained as a contaminated baseline and not submitted.
- **C, baseline+Q2:** only the Q2-bearing transcript rows are replaced by regenerated traces (1,282 repaired).
- **B, Q1+Q2:** every flagged row is replaced by a regenerated trace (8,788 repaired).

The variants form a dose-response sequence of 0, 1,282, and 8,788 repaired traces: A is a faithful contaminated baseline, C isolates the transcript decontamination, and B adds the option-discrimination repair on top. B and C are submitted; A is reported only as the baseline.

3.3. GRPO with a reference-guided reasoning judge

We further optimize variant C, the strongest SFT model on the development set (Section 5), with Group Relative Policy Optimization [2], initializing the policy from the SFT-C checkpoint and training on the held-out 30% partition. Each prompt produces a group of 8 sampled rollouts whose rewards are compared group-relative. We run GRPO under two reward designs. The full reward adds a reasoning-quality judge term,

$$R_{\text{judge}} = a + 0.5\, f + 0.2\, a\, \frac{s-1}{4}, \tag{1}$$

and the ablated reward drops that term, leaving accuracy and format alone,

$$R = a + 0.5\, f, \tag{2}$$

where $a \in \{0, 1\}$ marks a correct final answer, $f \in \{0, 1\}$ marks a well-formed completion (a reasoning body and exactly one trailing answer tag), and $s \in \{1, \ldots, 5\}$ is the judge's reasoning-quality score, active only when $a = 1$. Contrasting (1) with (2) isolates the judge's contribution; Section 5 shows the simpler reward (2) yields our strongest system.

The judge is Prometheus-7b-v2.0 [7]. It reads each rollout's reasoning against the row's reference CoT, which is the regenerated trace for flagged rows and the original Gemini CoT for clean rows, and never hears the audio, so it grades reasoning as text entailment rather than perception. Its intended role is to add variance: under a correctness-only reward an all-correct group yields no gradient, and a fabricated-but-correct rollout is reinforced identically to a grounded one. The judge is meant to break these ties toward specific, reference-consistent, per-option reasoning. We cap its weight at a small 0.2, against 1.0 for accuracy and 0.5 for format, so the policy optimizes the task rather than the judge; as Sections 5 and 6 discuss, this small cap may be why the judge moves accuracy so little. In both runs the development-optimal checkpoint is at 300 steps.

4. EXPERIMENTAL SETUP

Hardware and framework. All training and inference ran on the Illinois Campus Cluster and NCSA Delta using NVIDIA H200 and H100 GPUs, with full-parameter tuning in ms-swift [8] under DeepSpeed ZeRO-2 and bf16.

SFT. Each of the three variants trains the Qwen2.5-Omni-7B thinker with all parameters unfrozen (no LoRA) on six H200 GPUs, at a per-GPU batch size of 28 (global batch 168), for two epochs, with maximum sequence length 1024, learning rate 10^{-6}, a cosine schedule, and 5% warmup.

GRPO. Starting from the SFT-C checkpoint, we run on four H200 GPUs with vLLM-served colocated rollouts: 8 rollouts per prompt, sampling temperature 1.5 with top-k 4, one optimization pass per batch, KL coefficient $\beta = 10^{-3}$, learning rate 10^{-6} cosine, maximum completion length 768, and maximum sequence length 4096. We run 600 steps, evaluate the 300- and 600-step checkpoints on the development set, and submit the dev-optimal step-300 checkpoint.

Inference. All systems decode greedily with repetition penalty 1.05, a 512-token thinker budget, the KV cache enabled, FlashAttention-2, and batch size 16. Top-1 accuracy is scored by extracting the final answer tag and matching it to one of the four options.

Datasets. The official 1,607-item development set, DCASE2026-Task5-DevSet [9], draws a small portion from the MMAU [10], MMAR [11], and MMSU [12] benchmarks, with the majority newly constructed, human-annotated multiple-choice questions; it carries gold labels and is our sole basis for model selection and reported accuracy. The official ADQA-Bench evaluation set has 3,000 items whose gold labels are withheld during the challenge, so it yields predictions only.

5. RESULTS

Development-set ranking. Table 2 reports top-1 accuracy on the full 1,607-item development set, alongside the answer-tag parse rate. The parse rate is a format-compliance measure, not a correctness measure: it is the share of items on which the model emits a directly readable `<answer>` tag whose contents map to exactly

Table 2: Development-set top-1 accuracy and answer-tag parse rate (share of items emitting a directly-parseable answer that maps to one of the four options). "subm." is the submission index (label `Tathe_UIUC_task5_N`); rows marked *ref* are baselines and are not submitted.

system	subm.	acc. (%)	parse (%)
base Qwen2.5-Omni-7B	ref	53.08	86
SFT-A, baseline	ref	55.13	97
SFT-B, Q1+Q2	4	55.51	97
SFT-C, baseline+Q2	3	56.32	96
GRPO on C, with judge, 300	2	57.31	99
GRPO on C, no judge, 300	1	**58.43**	99

one of the four options, so the scorer can read the choice off without fallback heuristics, regardless of whether that choice is right. A trace can therefore parse cleanly yet be scored wrong, which is why parse rate and accuracy are reported as separate columns. Every fine-tuned variant improves over the un-tuned base model, and the best system is GRPO on variant C *without* the judge term, at 300 steps: 58.4% top-1, $+5.4$ points over base and $+2.1$ over its SFT-C starting point (McNemar vs. SFT-C $p = 0.057$, the strongest and nearly significant result we obtain).

The judge reward does not help. At 300 steps, dropping the Prometheus term (reward = accuracy + format) matches or slightly beats keeping it, 58.4% vs. 57.3%. The two are statistically tied (McNemar $p = 0.23$), but the point estimate moves *against* the judge and the no-judge run trains 2.4× faster, since the judge calls dominate step time. We therefore conclude that the reference-guided judge adds no measurable accuracy on this task, and our best system uses the simpler reward.

Over-training. Both GRPO runs peak at 300 steps and decay by 600 (no judge 58.4 to 57.1%, with judge 57.3 to 56.9%), so the decay is a property of prolonged GRPO rather than of the reward; we stop at 300 steps.

What the repair reliably fixes. The clearest and most robust effect is format compliance: the base model emits a clean answer tag on only 86% of items, whereas every fine-tuned system reaches 96 to 99%, recovering a large block of otherwise unparseable predictions. Among the SFT variants the three are statistically indistinguishable (paired McNemar, all $p > 0.2$), and only SFT-C significantly beats the base model ($p = 0.015$). The most-repaired variant (B) is not the strongest, and its much lower training loss reflects stylistic mimicry of the regenerated traces rather than better grounding, which is why model selection is done on held-out dev.

6. LIMITATIONS AND FUTURE WORK

Teacher reasoning quality. Our regenerated traces come mostly from the compact Gemma-4-E4B-it teacher, with a minority from the larger Qwen3-Omni-30B; both are smaller than the Gemini 3.1 Pro model that wrote the original CoT. Decontamination removes the harmful signal, but the accuracy gain is small (variant B lifts SFT only 55.13 to 55.51% over the baseline, and the lightly repaired C is the stronger model), so our teachers' reasoning, while grounded, may not match Gemini's quality; a stronger teacher pool is the natural next step.

Temporal traces. We could not tell from surface form whether a timestamp was genuinely heard or narrated from a transcript, so we left temporal evidence unflagged; some residual contamination on temporal items may remain.

The judge. The Prometheus reward did not improve accuracy and cost 2.4× the time. At its capped weight of 0.2 it may have been too weak to move the policy; revisiting it with a larger weight and a stronger or audio-aware grader is future work.

7. CONCLUSION

We treated DCASE 2026 Task 5 as a data-quality problem: a deterministic audit showed that most official Gemini 3.1 Pro CoT labels ignore the answer options or reason from a non-existent transcript, the behaviours the benchmark penalizes. Regenerating the flagged traces with a blind, audio-only teacher cascade under rejection sampling yields grounded, option-discriminating supervision, using only the provided data and a sub-10B model for the lightweight track. Our best system reaches 58.4% on dev with GRPO, the repaired SFT variants are statistically tied, and the most robust gain is answer-format compliance. A controlled ablation shows the judge adds no measurable accuracy at 2.4× the cost, so our strongest model uses a simple accuracy-and-format reward. We submit four systems: the two repaired SFT variants and the GRPO model with and without the judge.

8. ACKNOWLEDGMENT

This work used computing resources of the Illinois Campus Cluster and the NCSA Delta system at the University of Illinois Urbana-Champaign, running on NVIDIA H200 and H100 GPUs.

9. REFERENCES

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

[2] 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.

[3] 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 *Proc. Int. Conf. Learning Representations (ICLR)*, 2026.

[4] Google, "Gemma-4-E4B-it," https://huggingface.co/google/gemma-4-E4B-it, 2026, hugging Face model card. Accessed: June 2026.

[5] Qwen Team, "Qwen3-Omni technical report," *arXiv preprint arXiv:2509.17765*, 2025.

[6] ——, "Qwen2.5-Omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[7] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo, "Prometheus 2: An open source language model specialized in evaluating other language models," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[8] Y. Zhao *et al.*, "SWIFT: A scalable lightweight infrastructure for fine-tuning," in *Proc. AAAI Conf. Artificial Intelligence*, 2025.

[9] DCASE Community, "DCASE 2026 task 5 development set (DCASE2026-Task5-DevSet)," https://huggingface.co/datasets/Harland/DCASE2026-Task5-DevSet, 2026, accessed: June 2026.

[10] 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 *Proc. Int. Conf. Learning Representations (ICLR)*, 2025.

[11] Z. Ma, Y. Ma, Y. Zhu, C. Yang, Y.-W. Chao, R. Xu, W. Chen, Y. Chen, Z. Chen, *et al.*, "MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix," *arXiv preprint arXiv:2505.13032*, 2025.

[12] 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.