Long Listening Thoughts: Eliciting Open Auditory Reasoning with Deliberative Perception and Cognitive Refinement

Jaeyeon Kim[1] **Chao-Han Huck Yang**[2] **Luoyi Zhang**[1] **Chan-Jan Hsu**[1]
Fernando Ruiloba Portilla[1] **Jinchuan Tian**[1] **Shinji Watanabe**[1] **Carlos Busso**[1]
[1]Carnegie Mellon University [2]NVIDIA
jaeyeon2@andrew.cmu.edu huckiyang@ieee.org luoyiz@andrew.cmu.edu
chanjanh@andrew.cmu.edu feruiloba@gmail.com jinchuat@andrew.cmu.edu
shinjiw@ieee.org busso@cmu.edu

Abstract

We introduce Long Listening Thoughts (LLT), a framework for constructing post-training datasets that improves audio reasoning by pairing deliberate perceptual grounding with explicit cognitive behaviors. On audio understanding and reasoning benchmarks, large audio–language models (LALMs) are limited primarily by perceptual errors, yet existing post-training datasets rely on audio descriptions of limited diversity. Through a four-stage pipeline, LLT (1) builds multi-aspect dense audio descriptions spanning semantic, acoustic, and temporal information, (2) generates more challenging questions, (3) synthesizes incorrect reasoning traces, and (4) extends short reasoning traces into long ones via a thought-continuation strategy that elicits cognitive behaviors such as verification and backtracking. Using LLT, we synthesize 1.6M reasoning traces and train Qwen2.5-Omni-7B with a two-stage pipeline of supervised fine-tuning (SFT) followed by direct preference optimization (DPO). This improves accuracy on MMAU Test-mini (71.5 $\rightarrow$ 78.2), MMAR (56.7 $\rightarrow$ 63.7), and MMSU (60.6 $\rightarrow$ 66.0). Project page: https://jaeyeonkim99.github.io/llt

1 Introduction

Recent progress in large audio–language models (LALMs) has substantially advanced audio processing across a wide range of domains, including speech, sound events, music, or mixed acoustic scenes (Deshmukh et al., 2023; Radhakrishnan et al., 2023; Gong et al., 2024; Ghosh et al., 2024; Chu et al., 2024; Goel et al., 2025; Abouelenin et al., 2025; Tian et al., 2025, 2026). These models support diverse objectives ranging from traditional tasks such as automatic speech recognition (ASR) to tasks that require complex reasoning such as audio question answering (Yang et al., 2025b; Ma et al., 2026). To further enhance their reasoning capabilities, recent efforts have focused on post-training strategies such as constructing chain-of-thought (CoT) datasets (Zhifei et al., 2025; Kong et al., 2025; Wijngaard et al., 2025) for supervised fine-tuning (SFT), applying reinforcement learning (RL) techniques (Li et al., 2025; Rouditchenko et al., 2025; Wu et al., 2025; Fan et al., 2026), or combining both approaches by applying RL after SFT (Wen et al., 2025; He et al., 2025).

However, despite these advances, LALMs still exhibit notable limitations when evaluated on audio understanding and reasoning benchmarks. Recent benchmarks covering a broad range of audio reasoning tasks (Ma et al., 2025b; Sakshi et al., 2025; Wang et al., 2026a) show that perceptual errors account for the majority of failures, followed by reasoning errors. Moreover, LALMs are known to struggle with perceiving fine-grained details in audio, such as low-level acoustic attributes (Kim et al., 2025) and precise temporal structure (Wang et al., 2026b). Existing post-training approaches largely rely on audio descriptions generated by the LALMs themselves (Chen et al., 2025; Hu et al., 2024) or on ground-truth metadata (Lu et al., 2025; He et al., 2025) with limited diversity (e.g., coarse audio event labels in AudioSet (Gemmeke et al., 2017)) when constructing training data. This reliance constrains the model's ability to reason over richer and more deliberate perceptual cues that go beyond its original perceptual capacity. These observations suggest that post-training datasets guided by more deliberate, multi-aspect perceptual information could substantially improve audio reasoning performance.

In parallel, test-time scaling has been shown to be effective for complex reasoning tasks such as math and coding (Guo et al., 2025; Jaech et al., 2024). However, explicit strategies for leveraging longer reasoning traces in audio reasoning remain relatively underexplored. Prior work indicates that successful test-time scaling is closely related to structured cognitive behaviors such as verification

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37844–37876
October 24-29, 2026 ©2026 Association for Computational Linguistics

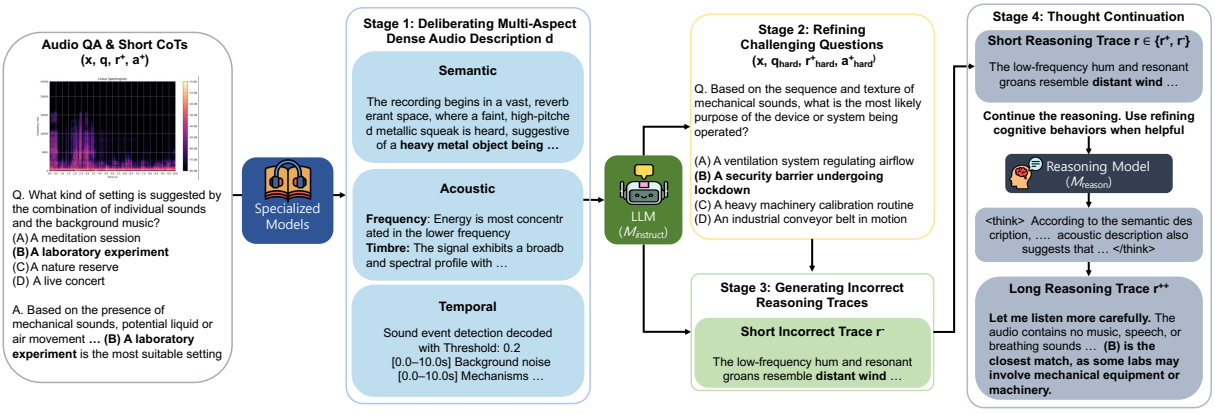


Figure 1: **Overview of the proposed *Long Listening Thoughts* pipeline. Stage 1** extracts multi-aspect audio descriptions (semantic, acoustic, temporal) from specialized models. **Stage 2** augments more challenging questions using an LLM. **Stage 3** generates incorrect reasoning traces. **Stage 4** extends traces via thought continuation to elicit cognitive behaviors.

and backtracking (Gandhi et al., 2025; Yeo et al., 2025) that enable models to explore alternative hypotheses and refine their reasoning. These behaviors have been further shown to be effective in perceptually oriented visual reasoning (Liao et al., 2025; Acuña et al., 2025), yet have not been systematically incorporated into audio reasoning frameworks.

Motivated by these observations, we propose Long Listening Thoughts (LLT), a new framework for constructing post-training datasets for audio reasoning that integrates deliberate multi-aspect perceptual grounding with explicit cognitive behaviors. Building upon existing audio CoT datasets, LLT introduces a scalable four-stage pipeline in which each stage targets a distinct factor underlying high-quality audio reasoning: **Stage 1** constructs multi-aspect dense audio descriptions to strengthen perceptual grounding, **Stage 2** augments more challenging questions that demand non-trivial reasoning, **Stage 3** synthesizes plausible incorrect reasoning traces that provide a corrective signal for reflection and correction, and **Stage 4** generates long reasoning traces with explicit cognitive behaviors. The overview is illustrated in Figure 1.

Using the LLT framework, we construct a post-training dataset comprising approximately 1.6M reasoning traces, supporting a two-stage training pipeline of SFT followed by Direct Preference Optimization (DPO) (Rafailov et al., 2023). We are sharing the dataset, trained models, and the full data construction pipeline to support reproducibility and future research. Through extensive analysis, we demonstrate the effectiveness of the proposed four-stage pipeline in generating perceptually grounded and cognitively structured reasoning traces. When training Qwen2.5-Omni-7B (Xu et al., 2025a) with LLT, we observe substantial improvements over the original model on audio reasoning benchmarks.

2 Related Work

Large Audio Language Models. Recent advances in LALMs have demonstrated impressive capabilities in processing diverse sound domains, including speech, sound events, and music (Deshmukh et al., 2023; Chu et al., 2024; Ding et al., 2025; Ghosh et al., 2024; Goel et al., 2025; Gong et al., 2024; Lu et al., 2025; Tian et al., 2025, 2026). Moreover, recent omni-modal models process audio inputs alongside visual inputs by integrating encoders from multiple modalities with an LLM. These approaches have also demonstrated strong performance in audio understanding (Xu et al., 2025a,b; Ye et al., 2025).

To systematically evaluate LALMs' capabilities, several benchmarks assess general audio understanding and reasoning across a wide range of tasks (Sakshi et al., 2025; Ma et al., 2025b; Wang et al., 2026a; Kumar et al., 2025). Analyses on these benchmarks show that audio reasoning remains a challenging task for LALMs, and that perceptual errors account for the majority of failures, followed by reasoning errors (Sakshi et al., 2025; Ma et al., 2025b). In this work, we propose a post-training framework that addresses this limitation by constructing reasoning traces grounded in richer, multi-aspect perceptual information and incorporating cognitive behaviors such as verification and backtracking, improving both perception and rea-

soning.

Audio Reasoning Models. To further enhance the reasoning capabilities of LALMs, recent works have proposed a variety of approaches targeting audio reasoning more explicitly. One direction leverages chain-of-thought reasoning traces, either through prompting (Ma et al., 2025a) or training. On the training side, prior work either folds CoT data into pre-training (Goel et al., 2025) or curates dedicated CoT datasets for SFT (Zhifei et al., 2025; Kong et al., 2025; Wijngaard et al., 2025). Across these efforts, however, the audio supervision is typically drawn from high-level event captions or metadata, providing only limited perceptual grounding. LLT departs from this by grounding traces in multi-aspect perceptual information, while extending prior CoT datasets through a scalable pipeline rather than constructing supervision from scratch.

Another direction applies online RL, typically Group-Relative Policy Optimization (GRPO) (Shao et al., 2024), on top of audio question-answering data. Several works apply RL directly to LALMs (Rouditchenko et al., 2025; Li et al., 2025; Zhao et al., 2025a; Wu et al., 2025), while others prime reasoning with SFT before training with GRPO (Wen et al., 2025; He et al., 2025). They aim to improve training curricula (Wen et al., 2025; He et al., 2025) or reward design (Fan et al., 2026; Wu et al., 2025) for GRPO. Rather than targeting the GRPO stage, LLT focuses on the upstream SFT stage to install perceptual grounding and cognitive behaviors. Specifically, we propose a post-training dataset and an SFT $\rightarrow$ DPO training pipeline for this purpose, while the resulting model remains compatible with GRPO for further capability gains.

3 Long Listening Thoughts

We propose Long Listening Thoughts (LLT), a framework for constructing post-training datasets of perceptually grounded and cognitively structured reasoning traces for audio reasoning. Rather than generating questions and reasoning traces from scratch (Zhifei et al., 2025; Goel et al., 2025; He et al., 2025), LLT enriches examples drawn from an existing audio CoT dataset, which improves scalability and keeps the resulting data compatible with prior datasets.

Notation. We write an example of the source audio CoT dataset as (x, q, r^+, a^+), where x is the audio, q the question, r^+ the correct short reasoning trace, and a^+ the correct answer. Stage 1 attaches a dense audio description d, extending the tuple to (x, q, r^+, a^+, d); when a component is irrelevant to a given step we write "$-$" in its place (e.g., $(-, q, r^+, a^+, d)$ for a text-only step that does not access the audio x). Superscripts mark correctness: r^- and a^- denote an *incorrect* short trace and answer, respectively (Stage 3). A *long* reasoning trace r^{++} extends a short trace with explicit cognitive behaviors that revisit and refine the initial reasoning (Stage 4; e.g., *"Wait, the initial reasoning incorrectly assumes. . . "*). Two language models drive the construction pipeline: an instruction-following LLM M_{instruct} and a reasoning LLM M_{reason}.

Overview. Given a source example, LLT applies the four stages illustrated in Figure 1: **Stage 1** builds the multi-aspect dense audio description d (Section 3.1); **Stage 2** rewrites the question into a more challenging one (Section 3.2); **Stage 3** synthesizes an incorrect short trace r^- (Section 3.3); and **Stage 4** continues a short trace $r \in \{r^+, r^-\}$ into a long trace r^{++} through thought continuation (Section 3.4). These stages yield three trace types per example: short correct r^+, *correct-to-refine* ($r^+ \rightarrow r^{++}$), and *incorrect-to-correct* ($r^- \rightarrow r^{++}$), where the arrow denotes thought continuation; their roles in training are described in Section 3.5.

3.1 Stage 1: Deliberating Multi-Aspect Dense Audio Description

As discussed in Section 1, perceptual errors are the dominant failure mode on audio reasoning benchmarks, motivating richer perceptual grounding. However, existing CoT datasets typically rely on audio metadata (He et al., 2025; Lu et al., 2025) or captions generated by LALMs themselves, such as Qwen2-Audio (Chu et al., 2024; Wen et al., 2025; Wu et al., 2025), which capture only limited aspects of the audio. One notable exception, AudSemThinker (Wijngaard et al., 2025), aggregates audio and visual analyses from multiple models, but still relies on Qwen2-Audio, audio captioning (Labbé et al., 2024), and event tagging (Chen et al., 2023) models that convey overlapping, high-level semantic information. Therefore, we propose to build a multi-aspect dense description for each audio input, providing a more holistic view of the auditory scene and richer cues for reasoning trace generation.

Figure 1 highlights the three complementary aspects of the audio information extracted in Stage 1:

(i) semantic information describing high-level audio events and their interpretations, (ii) acoustic information capturing attributes below the event level, such as duration, pitch, timbre, and amplitude, and (iii) temporal information providing precise timestamps and temporal structure. For semantic information, we use Qwen3-Omni-Captioner (Xu et al., 2025b), a recent specialized model to generate dense descriptions of audio events. For acoustic information, we directly leverage rich acoustic cues from spectrograms by prompting a vision-language model (VLM) to describe attributes such as frequency distribution, duration, spectral patterns, and reverberation. We find that providing explicit grid and axis information that the VLM can numerically ground, together with detailed interpretation guidelines (e.g., which acoustic attributes to attend to and illustrative examples), substantially improves the quality of spectrogram-based descriptions. Since LALMs typically lack a detailed understanding of low-level acoustic attributes, this spectrogram-driven approach yields more accurate and informative acoustic descriptions than relying solely on audio–language models (see Appendix D.1 for examples). We use Qwen3-VL-32B (Bai et al., 2025) as the VLM for spectrogram interpretation.

For temporal information, we employ specialized models to obtain accurate timestamps and temporal structure. We first select sub-tasks that collectively cover complementary aspects of temporal structure (speech, speaker turns, sound events, musical structure, and chord progression), and then choose strong-performing, open-source models for each. Specifically, Whisper-large-v3 (Radford et al., 2023) is used to generate timestamps for speech segments, while pyannote (Bredin, 2023) provides speaker diarization timestamps. In addition, we use a pretrained sound event detection model (Schmid et al., 2025) based on BEATs (Chen et al., 2023) to extract event boundaries. We apply multiple detection thresholds (0.2 and 0.5) for sound event detection in order to capture both prominent and subtle events. Finally, for audio clips detected to contain music, we additionally apply allin1 (Kim and Nam, 2023) music structure analysis and autochord chord recognition (Bayron, 2023) to obtain temporally aligned musical annotations. By concatenating all information, we form the dense audio description d. An example of the generated dense description is shown in Appendix D.2.

3.2 Stage 2: Refining Challenging Questions

With a more holistic view of the auditory scene, we can generate more challenging questions grounded in richer perceptual evidence. Given a source datapoint (x, q, r^+, a^+), we prompt an instruction-following large language model M_{instruct} to generate a harder datapoint $(x, q_{\text{hard}}, r^+_{\text{hard}}, a^+_{\text{hard}})$. This process is demonstrated in Stage 2 of Figure 1. Specifically, the model is instructed to retain the original question format and required skill set, while increasing difficulty by making the question more audio-aware or by requiring more complex reasoning. For instance, a content-recall question can be rewritten to demand closer listening (*"Summarize the speaker's advice"* → *"How does the speaker's intonation distinguish hopeless failure from greater success?"*) or deeper inference (*"What does the speaker claim about drama critics?"* → *"What does the other speaker's sarcastic reply reveal about her attitude?"*). Additionally, we verify the generated questions using M_{instruct} given the dense audio description d to ensure that (i) the question can be solved by listening to the audio, (ii) the final answer is correct, and (iii) the associated reasoning trace is logically valid. Empirically, when evaluated on 10k samples from AF-Think (Goel et al., 2025), Qwen2.5-Omni-7B's accuracy drops from 72.5% on the original questions to 63.0% on their hardened counterparts, confirming that Stage 2 produces meaningfully harder questions.

3.3 Stage 3: Generating Incorrect Reasoning Traces

To enhance reflective and corrective reasoning, we generate incorrect short traces r^- that contrast with the correct r^+. A natural source would be to directly sample from the LALM being trained by prompting it on questions and collecting its incorrect outputs, an approach explored in vision reasoning to build negative reasoning traces (Acuña et al., 2025; Liao et al., 2025). However, this strategy has several limitations. First, it does not consistently generate incorrect reasoning traces even when explicitly prompted to reason (Wu et al., 2025). Second, the resulting errors lack sufficient diversity to reflect the range of misperceptions that can arise in real-world scenarios. Third, traces sampled from a single model tie the negative supervision to that model's failure distribution, limiting the generalizability of the resulting dataset to other LALMs.

To address these limitations, we synthesize incorrect reasoning traces by prompting a language model with $(-, q, r^+, a^+, d)$. Specifically, we instruct the model to retain the length and structural format of the correct reasoning trace r^+, while generating a plausible but incorrect reasoning trace r^- that arises from misperceiving or misinterpreting fine-grained details in the dense audio description. This process encourages the generation of errors that are subtle, realistic, and grounded in perceptual ambiguity. This process is illustrated in Stage 3 of Figure 1. We generate r^- for both original questions and the challenging questions produced in Stage 2. The resulting incorrect reasoning traces are subsequently used in Stage 4 to generate long reasoning traces with explicit cognitive behaviors. An ablation comparing different sources of incorrect reasoning traces is provided in Table 3.

3.4 Stage 4: Thought Continuation

We augment short reasoning traces by appending continuations that exhibit explicit cognitive behaviors, building on the initial reasoning rather than replacing it. Specifically, we apply thought continuation to both correct traces r^+ from the source CoT dataset and Stage 2, and incorrect traces r^- from Stage 3, yielding two types of long reasoning traces: *correct-to-refine* traces $(r^+ \rightarrow r^{++})$ and *incorrect-to-correct* traces $(r^- \rightarrow r^{++})$.

Previous work in the vision domain adopts a *thought expansion* approach, distilling long reasoning traces from a large reasoning model M_{reason} by combining a short trace r with a cognitive trigger such as *"Wait"* or *"Hmm"* and prompting M_{reason} to complete the trace within the `<think></think>` block (Liao et al., 2025; Acuña et al., 2025). Applying this directly to audio is ineffective. Due to the length and multi-aspect nature of dense audio descriptions, M_{reason} struggles to determine which parts to attend to and produces excessively long, unfocused, or repetitive traces. This is undesirable for training LALMs, which should ground reasoning in the audio itself.

We instead adopt a *thought-continuation* approach. Given an initial trace $r \in \{r^+, r^-\}$, a cognitive trigger c, and the dense audio description d from Section 3.1, we provide $r \oplus c$ (where $\oplus$ denotes concatenation) in the prompt and have M_{reason} generate a continuation of the reasoning. The model deliberates on how to continue $r \oplus c$ inside the `<think></think>` block and produces the final continuation in the subsequent `<answer></answer>` block, naturally refining its own reasoning while following the original output format. The resulting long trace r^{++} keeps the initial reasoning r intact and appends a continuation that revisits and refines it, reaching the correct answer whether the prefix is r^+ (correct-to-refine) or r^- (incorrect-to-correct). Correctness is not guaranteed by construction; we enforce it through the verification step described below, which discards traces whose final answer is incorrect.

To encourage explicit cognitive behaviors within the continuation, we prompt M_{reason} to encode verification and backtracking, and use both general triggers (*Wait*, *Hmm*, *Alternatively*, *Let me check again*) and audio-aware triggers (*Let me listen again*, *Let me listen carefully*) as c to encourage grounding in the audio. This approach produces traces that are more concise and SFT-friendly than directly distilled thinking tokens; a comparison is shown in Appendix D.3.

In addition, we apply verification during long reasoning trace generation. Specifically, we prompt M_{instruct} with the generated continuation and $(-, q, a^+, d)$, and verify the following criteria: (i) whether the final answer is correct, (ii) whether the reasoning is logically valid, and (iii) whether the output format is valid (e.g., no references to the description text and proper use of the `<answer>` format). For criteria (i) and (ii), samples that fail verification are discarded. For criterion (iii), minor formatting errors are corrected with M_{instruct} during verification, while samples with severe format violations are discarded. We complement this automated verification with human inspection on a sampled subset (Appendix C).

3.5 Two-Stage Training

The LLT pipeline produces three types of reasoning traces per question: (x, q, r^+, a^+), $(x, q, r^+ \rightarrow r^{++}, a^+)$, and $(x, q, r^- \rightarrow r^{++}, a^+)$. While incorrect-to-correct traces $(r^- \rightarrow r^{++})$ from Stage 3 are essential for installing cognitive behaviors such as verification and backtracking, our analysis (Section 4.2) reveals that incorporating them directly into SFT hinders general reasoning performance, as the model struggles to distinguish intermediate r^- from r^+ under standard SFT supervision. To address this, we disentangle the role of each trace type across two training stages. We use only (x, q, r^+, a^+) and $(x, q, r^+ \rightarrow r^{++}, a^+)$ for SFT to strengthen perceptual grounding and refined reasoning, and reserve $(x, q, r^- \rightarrow r^{++}, a^+)$

for the subsequent DPO stage, where the preference structure naturally exploits the contrast between incorrect and corrected reasoning to induce cognitive behaviors. Unlike prior work (Liao et al., 2025) that places $(r^- \rightarrow r^{++})$ directly in SFT, our two-stage design avoids this conflict and yields both stronger performance and more pronounced cognitive behaviors (Section 4.2). Specifically, for each input (x, q) we construct three types of preference pairs that prefer correct over incorrect reasoning: *short-contrast* $(r^+ \succ r^-)$, *refine-contrast* $(r^+ \rightarrow r^{++} \succ r^-)$, and *correction-contrast* $(r^- \rightarrow r^{++} \succ r^-)$. The correction-contrast pair specifically teaches the model to prefer trajectories that recognize and correct initial errors over those that commit to them.

3.6 Implementation Details

We use AF-Think as the initial short CoT dataset, as it covers diverse audio domains and scenarios. For reproducibility, we exclude licensed data such as Music4all (Santana et al., 2020) and limit audio duration to 30 seconds for compatibility with general LALM architectures, yielding 338k audio clips, each with one original question–answer pair. Following AF-Think, we retain both multiple-choice and free-form questions. Stage 2 augments these with 316k harder questions, and Stages 3–4 produce the three reasoning-trace types, giving 1.6M traces in total that we release to support diverse post-training setups. For our two-stage pipeline, we use the 1.16M short-correct and correct-to-refine traces for SFT and sample 60k DPO preference pairs (20k per pair type; Section 3.5). Detailed dataset statistics are reported in Table 15.

We instantiate M_{instruct} as Qwen3-Next-80B-Instruct and M_{reason} as Qwen3-Next-80B-Thinking. All pipeline components, including the Stage-1 submodules, are open-source, ensuring full reproducibility. The prompts used at each stage are provided in Appendix E.

4 Experiments

4.1 Setup

Training. Following prior work (Wijngaard et al., 2025; Wu et al., 2025; He et al., 2025; Rouditchenko et al., 2025), we adopt Qwen2.5-Omni-7B as our base model, the backbone shared by most recent audio reasoning models. We apply our two-stage training as described in Section 3.5. While recent online-RL work for audio reasoning mainly focuses on improving GRPO training (Wu et al., 2025; Fan et al., 2026), we instead show that an LLT-trained model is itself a strong initialization that is readily compatible with online RL. We therefore additionally apply GRPO on AVQA (Yang et al., 2022), an audio question answering dataset widely used in prior audio RL work (Rouditchenko et al., 2025; Li et al., 2025). We randomly sample 8K questions and use a simple rule-based reward that assigns 1 if the answer inside the `<answer>` block is correct and 0 otherwise. More training details are given in Appendix F.

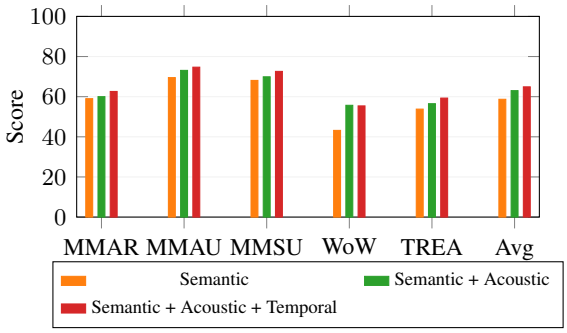


Figure 2: Accuracy of an LLM across different benchmarks using different types of audio descriptions. Semantic, Acoustic, and Temporal correspond to the respective components of the multi-aspect audio descriptions constructed in Stage 1 of LLT.

Evaluation Details. We evaluate on three widely used audio reasoning benchmarks: MMAU Test-mini (Sakshi et al., 2025) and MMAR (Ma et al., 2025b), which cover general audio understanding and reasoning across speech, sound, and music, and MMSU (Wang et al., 2026a), which focuses on speech perception and reasoning. More details are given in Appendix G.

4.2 Results

Does multi-aspect dense audio description benefit audio reasoning? To isolate the effect of description richness (Section 3.1) on reasoning, we feed audio descriptions with varying levels of information to a strong text LLM (Qwen3-32B (Yang et al., 2025a) in thinking mode, used only for this analysis) and have it answer the questions from the descriptions alone. We evaluate on general audio understanding and reasoning benchmarks (MMAU Test-mini, MMAR, MMSU) as well as specialized benchmarks targeting low-level acoustic attributes (WoW-Bench (Kim et al., 2025)) and temporal reasoning (TREA (Bhattacharya et al., 2025)).

Specifically, starting from the semantic description produced by Qwen3-Omni-Captioner, we progressively add acoustic and temporal information

Method	MMAU Test-mini	MMAR	MMSU
SFT with 50k QA Pairs			
AF-Think (r^+)	74.1	58.8	63.4
LLT ($r^+, r^+ \rightarrow r^{++}$)	**75.7**	**61.1**	**64.0**
LLT ($r^+, r^+ \rightarrow r^{++}, r^- \rightarrow r^{++}$)	74.6	60.2	63.3
SFT with Full Reasoning Traces			
AF-Think (r^+)	73.9	59.9	63.7
LLT ($r^+, r^+ \rightarrow r^{++}$)	**76.3**	**62.2**	**65.0**
LLT ($r^+, r^+ \rightarrow r^{++}, r^- \rightarrow r^{++}$)	75.8	61.4	64.1
Two-Stage Training			
LLT (SFT → DPO)	**78.2** (+1.9)	**63.7** (+1.5)	**66.0** (+1.0)

Table 1: Comparison across different training setups. Notation for reasoning traces follows Section 3. Parentheses indicate the trace types used for SFT. The last block reports the full two-stage pipeline (SFT → DPO). Best result in each block is bolded.

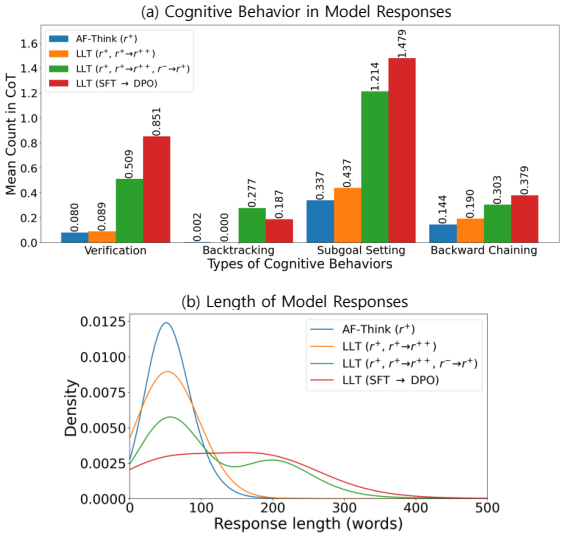


Figure 3: Cognitive analysis of model responses on MMAR. (a) Mean occurrence of different cognitive behaviors (b) Distribution of response length (number of words).

(Section 3.1) to form the full multi-aspect dense description. As shown in Figure 2, adding acoustic and temporal information yields consistent gains: each aspect benefits not only its corresponding specialized benchmark (WoW-Bench for acoustic, TREA for temporal) but also the general multi-task benchmarks. This result indicates that multi-aspect dense audio descriptions offer effective perceptual grounding and cues for generating high-quality reasoning traces in text-based LLMs.

Do incorrect reasoning traces benefit audio reasoning? Section 3.4 introduces correct-to-refine traces ($r^+ \rightarrow r^{++}$) and incorrect-to-correct traces ($r^- \rightarrow r^{++}$) in addition to the original short CoT r^+. While the latter is intended to encourage reflective reasoning, exposing the model to incorrect intermediate steps r^- may introduce noise during SFT. We compare three SFT settings under both a small-scale (50k QA) and a large-scale (full dataset) regime: r^+ only; r^+ with correct-to-refine ($r^+ \rightarrow r^{++}$); and all three types, additionally including incorrect-to-correct ($r^- \rightarrow r^{++}$).

As shown in Table 1, adding correct-to-refine traces achieves the best task accuracy, while additionally including incorrect-to-correct traces ($r^- \rightarrow r^{++}$) degrades performance, especially at small scale. We attribute this result to the model's difficulty in distinguishing intermediate r^- from r^+ under limited supervision, a gap that narrows with scale but is not fully resolved.

Analysis on model responses tells a different story. Quantifying cognitive behaviors on MMAR (Ma et al., 2025b) using the framework of Gandhi et al. (2025) and measuring response length (Figure 3), we find that models trained without ($r^- \rightarrow r^{++}$) exhibit substantially fewer cognitive behaviors and shorter traces. This result reveals a fundamental observation that ($r^- \rightarrow r^{++}$) is essential to install cognitive behaviors such as verification and backtracking, but directly including it in SFT compromises accuracy.

This finding motivates the two-stage design in Section 3.5, where ($r^- \rightarrow r^{++}$) is reserved for DPO rather than SFT. Using only 60k preference pairs from the full LLT trace pool, the resulting LLT SFT → DPO pipeline achieves both stronger task accuracy and more pronounced cognitive behaviors than any SFT setting alone, as shown in the last row in Table 1 and Figure 3.

We further verify that these cognitive behaviors translate into actual error recovery, not just surface-level reflection. We identify MMAR and MMAU Test-mini responses from the two-stage model that contain one of the cognitive triggers used in Stage 4 (e.g., *Let me listen again*, *Wait*), and compare the answer the model had reached *before* the cognitive trigger to its final answer after the subsequent continuation (Appendix B.1, Table 5). Across these triggered traces, 8.7% flip from incorrect to correct (true correction), while only 4.9% flip from correct to incorrect (over-correction), a statistically significant net correction of +3.8 percentage points (McNemar test (McNemar, 1947), $p<0.001$). Moreover, this triggered subset is intrinsically harder: the SFT-only model is substantially less accurate on triggered questions than on non-triggered ones (MMAR: 58.9% vs. 67.2%; MMAU: 72.6% vs. 81.4%; Fisher's exact test, both $p<0.01$). These results confirm that the two-stage model invokes

Method	MMAU Test-mini	MMAR	MMSU
Large Audio Language Models			
Qwen2-Audio-Instruct (2024)	59.6	30.0	53.3
Kimi-Audio (2025)	68.2	57.6	59.3
Audio Flamingo 3 (2025)	73.3	58.5	61.4
Audio Reasoning Models			
SFT-based			
Audio-Reasoner (2025)	67.7	36.8	49.2
AudSemThinker†‡ (2025)	71.2	50.0	60.2
GRPO on AVQA-Inclusive sets			
R1-AQA (2025)	68.9	50.8	61.6
Ke-Omni-R† (2025a)	74.6	60.9	62.1
Omni-R1† (2025)	77.0	63.4	–
Omni-R1 (AVQA)†	75.8	61.2	–
CESAR† (2026)	77.1	62.7	**64.2**
Audio-Thinker† (2025)	**77.7**	**65.3**	–
Our Methods			
Qwen2.5-Omni-7B	71.5	56.7	60.6
+ LLT (SFT → DPO)	78.2	63.7	**66.0**
+ LLT + AVQA GRPO	**78.7**	**64.9**	65.3

Table 2: Performance comparison across different models. Omni-R1 (AVQA) indicates the variant trained on the naive AVQA dataset. For clarity, we highlight in bold the best-performing existing method and the best-performing model among our approaches. † denotes methods built on the same Qwen2.5-Omni-7B backbone as ours; ‡ denotes results reproduced by us.

Setting	MMAU Test-mini	MMAR	MMSU
Qwen2.5-Omni baseline	71.5	56.7	60.6
Stage 3: Source of Incorrect Traces			
SFT on 50k QA Pairs			
From base LALM	72.9	58.3	**64.4**
From M_{instruct}	**74.6**	**60.2**	63.3
Two-Stage Training			
From base LALM	77.9	**63.7**	65.8
From M_{instruct}	**78.2**	**63.7**	**66.0**
DPO Pair Composition			
LLT (SFT → DPO, full)	78.2	**63.7**	**66.0**
$-\ (r^- \rightarrow r^{++} \succ r^-)$	78.3	62.2	64.4
$+\ (r^+ \succ r^+ \rightarrow r^{++})$	**78.4**	62.4	65.0
Generalization to AudioMCQ (He et al., 2025)			
r^+ only	73.2	57.7	63.8
$r^+, r^+ \rightarrow r^{++}$	**73.8**	**61.1**	**64.3**

Table 3: Ablation studies on the LLT pipeline. Each block isolates one design choice: the source of incorrect traces (Stage 3), DPO pair composition, and generalization to a different source dataset. The best result in each block is in bold.

cognitive behaviors selectively on harder questions, and that these in-trace revisions recover the correct answer more often than they introduce errors.

Comparison to baselines. The results are given in Table 2. Our two-stage LLT pipeline achieves substantial improvements over the Qwen2.5-Omni-7B baseline, with consistent gains across MMAU Test-mini, MMAR, and MMSU. Applying GRPO on AVQA (Yang et al., 2022), we observe further improvements in MMAR and MMAU Test-mini. We additionally compare against LALMs of similar size (7B) and audio reasoning models using SFT or GRPO on AVQA-inclusive datasets. Our approach outperforms most of them while requiring only SFT and DPO, and performs comparably to those that use LLM-based (Wu et al., 2025) or complex process rewards (Fan et al., 2026). Moreover, compared to Omni-R1 (AVQA) (Rouditchenko et al., 2025) and Ke-Omni-R (Zhao et al., 2025a), which share the same backbone of Qwen2.5-Omni-7B and outcome-based reward, LLT initialization yields stronger GRPO performance, indicating a stronger starting point for online RL.

Ablation on Incorrect Reasoning Traces (Stage 3). As described in Section 3.3, we synthesize incorrect reasoning traces using M_{instruct} rather than sampling them from the base LALM. A natural alternative is to prompt the base LALM on the source questions and retain outputs whose final answers are incorrect. We compare these two sources under two settings: (i) a small-scale SFT setting that fine-tunes on 50k QA pairs using all three trace types (r^+, $r^+ \rightarrow r^{++}$, and $r^- \rightarrow r^{++}$), and (ii) our two-stage setting, where DPO is applied on top of the full-trace SFT model using preference pairs whose r^- comes from the respective source. As shown in the second block of Table 3, synthesized traces generally outperform naturally occurring ones, indicating that the controlled, perceptually grounded errors produced by our pipeline provide a more effective corrective signal. More detailed analysis is given in Appendix B.3.

Ablation on DPO Pair Selection. The two-stage pipeline (Section 3.5) combines three types of preference pairs. To verify whether the incorrect-to-correct pair $(r^- \rightarrow r^{++} \succ r^-)$ genuinely contributes to the observed improvements, we ablate two variants. First, we remove the $(r^- \rightarrow r^{++} \succ r^-)$ pair, leaving only $(r^+ \succ r^-)$ and $(r^+ \rightarrow r^{++} \succ r^-)$, and increase each remaining pair type to 30k to match the original total scale. Second, we add a compactness-preference pair $(r^+ \succ r^+ \rightarrow r^{++})$ at 20k, following prior work (Liao et al., 2025; Acuña et al., 2025) that uses such pairs to prefer shorter, compact reasoning when the model already reaches the correct answer. Results are presented in the third block of Table 3.

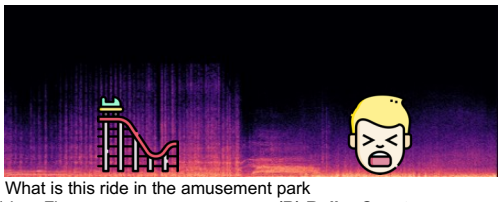
Method	MMAU Test-mini	MMAR	MMSU
Qwen2-Audio-Instruct	59.6	30.0	53.3
+ LLT SFT	**67.8**	**53.8**	**59.0**
Qwen3-Omni-Instruct	77.5	66.9	69.0
+ LLT SFT	**79.6**	**67.9**	**71.6**

Table 4: Generalization of LLT to additional LALM backbones.

Including the $(r^- \rightarrow r^{++} \succ r^-)$ pair performs comparably on MMAU Test-mini while outperforming both variants on MMAR and MMSU, and adding the compactness pair $(r^+ \succ r^+ \rightarrow r^{++})$ does not help. This confirms that our DPO pair selection is well-suited for the two-stage pipeline.

Applying LLT Pipeline to a Different Dataset. To investigate whether the four-stage LLT pipeline generalizes beyond AF-Think, we apply the same pipeline to a different source dataset. Specifically, we randomly sample 50k questions from AudioMCQ (He et al., 2025) and apply Stage 1 and Stage 4 to construct corresponding r^{++} traces for each r^+. We then compare two training settings: (i) using the original r^+ alone, and (ii) using both r^+ and $r^+ \rightarrow r^{++}$, following the SFT setup of our two-stage training pipeline. Results are presented in the fourth block of Table 3. Adding the $r^+ \rightarrow r^{++}$ traces consistently improves performance over the r^+-only baseline, demonstrating that our pipeline generalizes across different source datasets.

Applying LLT Dataset to Additional Models. To examine whether LLT generalizes beyond the primary backbone of Qwen2.5-Omni-7B, we fine-tune additional LALMs on a subset of the LLT dataset consisting of r^+ and $r^+ \rightarrow r^{++}$ reasoning traces corresponding to SFT examples derived from 50K QA pairs. Specifically, we fine-tune Qwen2-Audio-Instruct (Chu et al., 2024), which has been widely used in prior audio reasoning work (Ma et al., 2025a; Zhifei et al., 2025; Li et al., 2025), and Qwen3-Omni-Instruct (Xu et al., 2025b), a more recent and stronger LALM. For Qwen3-Omni-Instruct, we apply LoRA (Hu et al., 2022) with rank 64 and alpha 128 due to compute constraints. The results are shown in Table 4. Across benchmarks, both models trained with LLT show consistent improvements over their base counterparts, suggesting that LLT generalizes to other LALMs, including a stronger backbone such as Qwen3-Omni-Instruct. Additional experimental results are provided in Appendix B.



Q. What is this ride in the amusement park
(A) Log Flume **(B) Roller Coaster**
(C) Pirate Ship (D) Ferris Wheel

Qwen2.5-Omni-7B + AF-Think

<think> **The sounds of a motor starting, water splashing, and a scream suggest a ride involving water and speed.** A roller coaster or ferris wheel is less likely to have water splashing sounds. A pirate ship is less likely to have a motor starting sound. Therefore, the most plausible answer is**: (A) Log Flume.** </think>

<answer> **(A)** </answer>

Qwen2.5-Omni-7B + LLT (SFT → DPO)

<think> The presence of screams, **water splashes, and a boat engine suggests a water-based ride**. The fearful tone of the speaker further points towards a thrilling experience. Therefore, the most plausible option is: (A) Log Flume.

Let me listen more carefully. The audio features a deep, resonant roar followed by a high-pitched scream, with metallic clanking and grinding sounds. **The mechanical noises are consistent with a roller coaster's chain lift mechanism and track movement, not water-based elements.** The absence of water splashes or boat engine sounds rules out a log flume. The sudden, intense scream and rapid mechanical transitions align with a roller coaster's drop or inversion, where riders experience sharp drops and high-speed motion. The reverberant, enclosed environment described matches a roller coaster tunnel or station, not a pirate ship or Ferris wheel. **The initial assumption of water sounds was incorrect; the metallic and mechanical nature of the audio is definitive for a roller coaster. </think>**

<answer> **(B)** </answer>

Figure 4: Qualitative comparison of responses from Qwen2.5-Omni-7B trained with AF-Think and with the proposed LLT on the same question.

Qualitative Examples. As shown in Figure 4, compared to the model trained with AF-Think, which produces an incorrect answer, the model trained with LLT initially makes a mistake but subsequently revisits the audio, applies more deliberate perceptual analysis, and corrects its reasoning to arrive at the correct answer. More examples are available on the project page[1].

5 Conclusions

We introduce Long Listening Thoughts (LLT), a framework for constructing post-training datasets for audio reasoning that explicitly targets perceptual grounding and cognitive refinement. By integrating multi-aspect dense audio descriptions with a thought-continuation strategy, LLT produces 1.6M reasoning traces that encourage reflective and corrective reasoning behaviors. Extensive experiments show that models trained with LLT achieve improved audio reasoning across multiple benchmarks. We are sharing the dataset, models, and full data construction pipeline to support reproducibility and facilitate future research.

[1] https://jaeyeonkim99.github.io/llt

Limitations

While LLT consistently improves audio reasoning, several limitations should be acknowledged. First, LLT enriches an existing audio CoT dataset rather than collecting new audio, and our study focuses on clips of up to 30 seconds. This design keeps the pipeline scalable and compatible with prior datasets, but the resulting gains may not transfer directly to substantially longer recordings or to domains underrepresented in the source data. Extending the pipeline along these axes is a natural direction for future work.

Second, following prior audio reasoning data-construction pipelines (Wijngaard et al., 2025; He et al., 2025), LLT relies largely on model-based generation and verification. We complement this with human inspection on a sampled subset (Appendix C), but such automated verification cannot fully guarantee the correctness and quality of every trace at scale.

Finally, the fine-grained perceptual capabilities that LLT strengthens, such as acoustic attribute recognition, temporal localization, and speaker diarization, are intended to support reasoning, but in principle could also be applied to infer sensitive information from audio. To mitigate this, we build only on publicly available research datasets.

Acknowledgments

This material is based upon work supported in whole or in part with funding from the Department of Defense. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DoD and/or any agency or entity of the United States Government.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

David Acuña, Chao-Han Huck Yang, Yuntian Deng, Jaehun Jung, Ximing Lu, Prithviraj Ammanabrolu, Hyunwoo Kim, Yuan-Hong Liao, and Yejin Choi. 2025. Long grounded thoughts: Distilling compositional visual reasoning chains at scale. *arXiv preprint arXiv:2511.05705*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Christopher J. Bayron. 2023. autochord: Automatic chord estimation. https://github.com/cjbayron/autochord. Accessed: 2026-01-28.

Debarpan Bhattacharya, Apoorva Kulkarni, and Sriram Ganapathy. 2025. Benchmarking and confidence evaluation of LALMs for temporal reasoning. In *INTERSPEECH*.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *INTERSPEECH*.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and EngSiong Chng. 2025. Audio large language models can be descriptive speech quality evaluators. In *ICLR*.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *ICML*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*.

Jiajun Fan, Roger Ren, Jingyuan Li, Rahul Pandey, Prashanth Gurunath Shivakumar, Ivan Bulyko, Ankur Gandhe, Ge Liu, and Yile Gu. 2026. Incentivizing consistent, effective and scalable reasoning capability in audio llms via reasoning process rewards. In *ICLR*.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. In *CoLM*.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *ICASSP*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *EMNLP*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2025. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. In *NeurIPS*.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2024. Listen, think, and understand. In *ICLR*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, and 1 others. 2025. Measuring audio's impact on correctness: Audio-Contribution-Aware Post-Training of large audio language models. *arXiv preprint arXiv:2509.21060*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *Iclr*.

Yuchen Hu, CHEN CHEN, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and EngSiong Chng. 2024. Large language models are efficient learners of noise-robust speech recognition. In *ICLR*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*.

Jaeyeon Kim, Heeseung Yun, Sang Hoon Woo, Chao-Han Huck Yang, and Gunhee Kim. 2025. WoW-Bench: Evaluating fine-grained acoustic perception in audio-language models via marine mammal vocalizations. *arXiv preprint arXiv:2508.20976*.

Taejun Kim and Juhan Nam. 2023. All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio. In *WASPAA*.

Zhifeng Kong, Arushi Goel, Joao Felipe Santos, Sreyan Ghosh, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2025. Audio Flamingo Sound-CoT technical report: Improving chain-of-thought reasoning in sound understanding. *arXiv preprint arXiv:2508.11818*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, and 1 others. 2025. MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*.

Étienne Labbé, Thomas Pellegrini, and Julien Pinquier. 2024. Conette: An efficient audio captioning system leveraging multiple datasets with task embedding. *IEEE/ACM TASLP*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.

Yuan-Hong Liao, Sven Elflein, Liu He, Laura Leal-Taixé, Yejin Choi, Sanja Fidler, and David Acuña. 2025. LongPerceptualThoughts: Distilling system-2 reasoning for system-1 perception.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, and 1 others. 2025. DeSTA2.5-Audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *arXiv preprint arXiv:2507.02768*.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025a. Audio-CoT: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, and 1 others. 2025b. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *NeurIPS*.

Ziyang Ma, Ruiyang Xu, Yinghao Ma, Chao-Han Huck Yang, Bohan Li, Jaeyeon Kim, Jin Xu, Jinyu Li, Carlos Busso, Kai Yu, and 1 others. 2026. The interspeech 2026 audio reasoning challenge: Evaluating reasoning process quality for audio reasoning models and agents. *arXiv preprint arXiv:2602.14224*.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. In *NeurIPS*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A Kiani, David Gomez-Cabrero, and Jesper N Tegner. 2023. Whispering LLaMA: A cross-modal generative error correction framework for speech recognition. In *EMNLP*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*.

Harshit Rajgarhia, Shuubham Ojha, Asif Shaik, Akhil Pothanapalli, Rachuri Lokesh, Abhishek Mukherji, and Prasanna Desikan. 2026. Medmosaic: A challenging large scale benchmark of diverse medical audio. In *ICML*.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. 2025. Omni-R1: Do you really need audio to fine-tune your audio llm? *arXiv preprint arXiv:2505.09439*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *ICLR*.

Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Valéria Delisandra Feltrim, Marcos Aurélio Domingues, and 1 others. 2020. Music4all: A new music database and its applications. In *IWSSIP*, pages 399–404.

Florian Schmid, Tobias Morocutti, Francesco Foscarin, Jan Schlüter, Paul Primus, and Gerhard Widmer. 2025. Effective pre-training of audio transformers for sound event detection. In *ICASSP*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and 1 others. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jinchuan Tian, William Chen, Yifan Peng, Jiatong Shi, Siddhant Arora, Shikhar Bharadwaj, Takashi Maekaku, Yusuke Shinohara, Keita Goto, Xiang Yue, and 1 others. 2025. OpusLM: A family of open unified speech language models. In *Interspeech*.

Jinchuan Tian, Haoran Wang, Bo-Hao Su, Chien-yu Huang, Qingzheng Wang, Jiatong Shi, William Chen, Xun Gong, Siddhant Arora, Chin-Jou Li, and 1 others. 2026. Bagpiper: Solving open-ended audio tasks via rich captions. *arXiv preprint arXiv:2602.05220*.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2026a. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. *ICLR*.

Hualei Wang, Yiming Li, Shuo Ma, Hong Liu, and Xiangdong Wang. 2026b. Listening between the frames: Bridging temporal gaps in large audio-language models. In *AAAI*.

Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. 2025. SARI: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*.

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. 2025. AudSemThinker: Enhancing audio-language models through reasoning over semantics of sound. In *NeurIPS*.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. 2025. Audio-Thinker: Guiding audio language model when to think and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025b. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chao-Han Huck Yang, Sreyan Ghosh, Qing Wang, Jaeyeon Kim, Hengyi Hong, Sonal Kumar, Guirui Zhong, Zhifeng Kong, S Sakshi, Vaibhavi Lokegaonkar, and 1 others. 2025b. Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge. *arXiv preprint arXiv:2505.07365*.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. AVQA: A dataset for audio-visual question answering on videos. In *ACM MM*.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, and 1 others. 2025. OmniVinci: Enhancing architecture and data

for omni-modal understanding LLM. *arXiv preprint arXiv:2510.15870*.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. In *ICML*.

Shuaijiang Zhao, Tingwei Guo, Cheng Wen, Bajian Xiang, Wei Zou, and Xiangang Li. 2025a. Ke-omni-r: Achieving advanced audio reasoning with a concise 50-words think process. https://github.com/shuaijiang/Ke-Omni-R.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2025b. SWIFT: A scalable lightweight infrastructure for fine-tuning.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *ACL Demo*.

Xie Zhifei, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-Reasoner: Improving reasoning capability in large audio language models. In *EMNLP*.

A Extended Related Work

Recent large reasoning models (Jaech et al., 2024; Guo et al., 2025) have demonstrated impressive performance on complex reasoning tasks such as mathematics and code generation by leveraging increased computation at inference time and producing long reasoning traces. Similar trends have also been observed in vision-centric reasoning tasks (Liao et al., 2025; Acuña et al., 2025), where longer reasoning traces enable models to refine predictions through extended reasoning. In contrast, explicit strategies for test-time scaling in the audio reasoning domain remain largely underexplored.

Prior work indicates that effective test-time scaling is closely tied to structured cognitive behaviors, such as verification and backtracking (Gandhi et al., 2025; Yeo et al., 2025). In the visual domain, approaches such as Long Perceptual Thoughts (Liao et al., 2025) and Long Grounded Thoughts (Acuña et al., 2025) attempt to encode these cognitive behaviors by distilling long reasoning traces from large reasoning models, leading to improved perceptual and grounded reasoning. Inspired by these efforts, we propose Long Listening Thoughts (LLT), which explicitly encodes cognitive behaviors for auditory reasoning. Unlike prior visual approaches that rely on detailed descriptions derived from relatively small-scale image datasets, LLT is designed for scalability by building upon existing large-scale audio CoT datasets. In addition, LLT incorporates a principled pipeline for constructing multi-aspect audio descriptions spanning semantic, acoustic, and temporal information. Finally, rather than directly distilling raw long reasoning traces, we adopt a thought-continuation strategy that more effectively teaches cognitive behaviors while producing concise and structured reasoning traces suitable for audio reasoning.

B Additional Experimental Results

Pre → Post	**# Responses**	**%**
correct → correct	668	59.0
incorrect → incorrect	309	27.3
incorrect → correct *(correction)*	99	8.7
correct → incorrect *(over-correction)*	56	4.9
Net correction	**+43**	**+3.8**

Table 5: Pre-trigger vs. post-trigger answer transitions on MMAR and MMAU responses from the two-stage model (LLT SFT → DPO) that contain a Stage 4 cognitive trigger, over the 1,132 responses with a determinable pre-trigger answer. The net correction is statistically significant (McNemar test, $p<0.001$).

B.1 Cognitive Trigger Analysis

To verify that cognitive behaviors learned through the two-stage pipeline lead to actual error recovery, we analyze MMAR and MMAU Test-mini responses from the LLT SFT → DPO model that contain a Stage 4 cognitive trigger (e.g., *Let me listen again*, *Wait*, *Let me check again*). For each such response, we split the trace at its first cognitive trigger and prompt GPT-4.1-mini to identify the answer the model had committed to immediately before the trigger; responses for which no committed pre-trigger answer can be determined are excluded. We then compare this pre-trigger answer with the final answer after the continuation, pooling both benchmarks for a reliable significance test. Table 5 reports the resulting pre-trigger → post-trigger transitions over 1,132 determined responses. True corrections (99 cases, 8.7%) outnumber over-corrections (56 cases, 4.9%), yielding a statistically significant net gain of $+3.8$ percentage points on the triggered subset (McNemar test (McNemar, 1947), $p<0.001$). The effect holds on each benchmark individually (MMAR: $+4.0$ pp, $p=0.019$; MMAU: $+3.6$ pp, $p=0.019$).

Setting	MMAU Test-mini	MMAR	MMSU
Thought expansion	75.6	60.5	63.8
Thought continuation	**75.7**	**61.1**	**64.0**

Table 6: Ablation on long-trace construction (Stage 4). The best result is in bold.

Trace (role in DPO)	Base	SFT ckpt (DPO init)
r^+ (chosen)	2.173	0.715
$r^+ \rightarrow r^{++}$ (chosen)	2.153	0.831
$r^- \rightarrow r^{++}$, synthetic (chosen)	2.348	1.428
$r^- \rightarrow r^{++}$, model-sampled (chosen)	1.774	1.282
r^-, synthetic (rejected)	2.510	1.948
r^-, model-sampled (rejected)	0.767	1.615

Table 7: Per-token NLL of each trace type (lower = more likely under the model).

B.2 Ablation on Thought Continuation (Stage 4)

We compare the thought expansion approach used in prior work (Liao et al., 2025; Acuña et al., 2025), which distills long reasoning traces from large reasoning models, with our proposed thought continuation strategy. Both are evaluated under an SFT setting trained with r^+ and $r^+ \rightarrow r^{++}$, matching the SFT stage of our two-stage pipeline. The results are in Table 6. Thought continuation consistently outperforms thought expansion, demonstrating that generating refined reasoning traces after the initial thinking phase is more effective than directly distilling raw long reasoning traces for audio reasoning.

B.3 Analysis on Synthetic Incorrect Traces

We further analyze why the incorrect traces synthesized in Stage 3 (Section 3.3) are more effective than negatives sampled from the base LALM, considering two angles: coverage and likelihood.

Coverage. Model-sampled negatives are only obtainable when the base LALM both answers incorrectly and produces a reasoning trace. From 50k source questions, after running the full four-stage pipeline including verification, a single sampling run from the base LALM yields 4.1k $r^- \rightarrow r^{++}$ traces, whereas synthesis with M_{instruct} yields 31.4k. A second sampling run from the base LALM adds 4.3k more, but 1.8k of these are duplicate questions that are already covered. Model-based sampling is therefore difficult to scale, and limits the diversity of the errors it can express.

Likelihood analysis. To understand the mechanism beyond coverage, we measure the per-token negative log-likelihood (NLL) of each trace type under the base LALM (Qwen2.5-Omni-7B) and under the SFT checkpoint that serves as the reference policy for DPO, on 4k paired samples.

Preference pair	Synthetic	Model-sampled
$r^+ \succ r^-$	1.233	0.900
$r^+ \rightarrow r^{++} \succ r^-$	1.117	0.784
$r^- \rightarrow r^{++} \succ r^-$	0.520	0.333

Table 8: Separation between chosen and rejected responses under the DPO reference policy.

Model	Overall	≤30s (n=1720)	>30s (n=2862)
Qwen2.5-Omni-7B (reported)	52.2	–	–
Qwen2.5-Omni-7B (reproduced)	50.1	42.6	57.1
+ LLT (SFT → DPO)	**55.2**	**48.7**	**59.8**

Table 9: Results on MMAU-Pro.

The results are given in Table 7, suggesting different explanations for the two training stages. For SFT, which starts from the base model, model-sampled r^- is already highly likely (0.767), far below r^+ (2.173). Note that r^+ comes from an existing CoT dataset and is thus unfamiliar to the model. Including such traces therefore exposes the model to little new information and may reinforce its existing error patterns, whereas synthetic r^- (2.510) is unfamiliar to the base model.

For DPO, one might expect synthetic negatives to lie outside the model's distribution, but in practice the gap is small. Since r^- is generated to follow the format and length of r^+, its NLL drops substantially after SFT (2.510 → 1.948), while model-sampled r^- moves in the opposite direction (0.767 → 1.615). The gap between the two pools narrows from 1.743 to 0.333 under the reference policy. Synthetic negatives are also more clearly separated from the chosen responses under the reference policy, as measured by NLL(rejected)−NLL(chosen), across all three preference pair types used in our DPO stage (Table 8). This separation alone does not explain the performance difference, but together with the coverage gap above it characterizes how the two pools differ as preference data.

B.4 Generalization to Long-Form Audio

To examine whether our improvements transfer to longer audio, we evaluate on MMAU-Pro (Kumar et al., 2025), which includes recordings of up to 10 minutes with an average duration of about 2 minutes. Table 9 reports the results. Despite training exclusively on clips of up to 30 seconds, the gains

are retained on the >30s subset, indicating that the improved audio reasoning from LLT transfers, at least to some extent, to longer recordings.

B.5 Generalization to Specialized Audio Domains

To investigate whether these improvements transfer beyond general audio benchmarks, we additionally evaluate on two specialized domains. For the medical domain, we use the publicly released subset of MedMosaic (Rajgarhia et al., 2026) (3,982 multiple-choice instances), reported in Table 10. For bioacoustics, we use WoW-Bench (Kim et al., 2025), which is based on marine mammal vocalizations, reported in Table 11. In both cases the trained models improve overall performance, showing that LLT helps, at least to some extent, not only on general benchmarks but also in specialized domains.

B.6 Analysis on the Source of Gains

To further investigate the source of the gains, we evaluate our models under two conditions: removing the audio entirely, and replacing it with Gaussian noise, following prior work (Sakshi et al., 2025; Rouditchenko et al., 2025; He et al., 2025). All conditions are evaluated under an identical protocol for both models to isolate the effect of the audio input. Table 12 reports the results.

The results show two effects. First, without any audio, our model still outperforms the base model, indicating that text-side reasoning itself improves. Second, when the audio is replaced with Gaussian noise, our model degrades by a larger margin than the base model and even falls below its own no-audio score, suggesting that it actively grounds its reasoning in the acoustic evidence and is therefore misled when that evidence is corrupted. Together, these results indicate that the gains reflect both improved reasoning and strengthened perceptual grounding.

C Human Inspection of Reasoning Traces

To complement the automated LLM-based verification described in Section 3.4, we conduct a human inspection of the final reasoning trace pairs in our dataset. This inspection assesses data fidelity along multiple axes including audio grounding, textual artifacts, and reasoning quality.

C.1 Evaluation Protocol

We randomly sample 96 trace pairs from the final dataset: 48 *Type A* pairs $(r^+, r^+ \rightarrow r^{++})$ and 48 *Type B* pairs $(r^-, r^- \rightarrow r^{++})$. Twelve annotators, all graduate students with prior research experience in audio or multimodal machine learning, participated on a voluntary basis, each evaluating 16 pairs. Each pair was independently evaluated by two annotators, yielding 192 trace-level annotations.

Annotators were instructed to listen to the audio before evaluating each pair. The interface displayed the audio clip, the question with the correct answer marked, the initial reasoning trace (Left), and the continuation (Right). The textual audio description was not shown to annotators. The full guidelines and the labeling interface are shown in Figure 5 and Figure 6, respectively. The evaluation proceeds in two phases.

Question validity. Annotators first judge whether the question is valid: *Is the question answerable given the audio?* Yes if the question can be reasonably answered after listening, possibly with some external knowledge to interpret. No if the question depends on audio events or attributes that are not present at all, making it impossible to answer in any grounded way or to meaningfully evaluate the reasoning. If the question is judged invalid, the trace-level evaluation is skipped for that pair.

Per-trace evaluation. For pairs with valid questions, annotators independently evaluate the left and right traces along binary (Yes/No) criteria:

- **Hallucination**: Does the trace mention audio events or attributes with no audible basis in the audio (e.g., claiming "a dog barks" when no dog-like sound is present)?
- **Textual artifact**: Does the trace explicitly reference information not derivable from listening to the audio alone, suggesting it was borrowed from an external textual description (e.g., "the description mentions…," "according to the metadata…," "the caption says…")?
- **Reasoning valid**: Independent of the correct answer, is the reasoning logically coherent and plausible step-by-step?

For Type B pairs, the left trace (r^-) is by design incorrect, so *hallucination* and *reasoning valid* are

Model	Weighted Avg	Long-Form	Cough	Heart	Lung	Speech	Speech + Sound	Multi-Turn	Voice QA
Qwen2.5-Omni-7B	47.5	**54.7**	44.4	38.0	36.1	**55.6**	46.1	**100.0**	**56.1**
+ LLT (SFT → DPO)	**51.3**	45.3	**51.1**	**45.8**	**44.3**	55.3	**52.8**	88.9	47.8

Table 10: Results on the publicly released subset of MedMosaic.

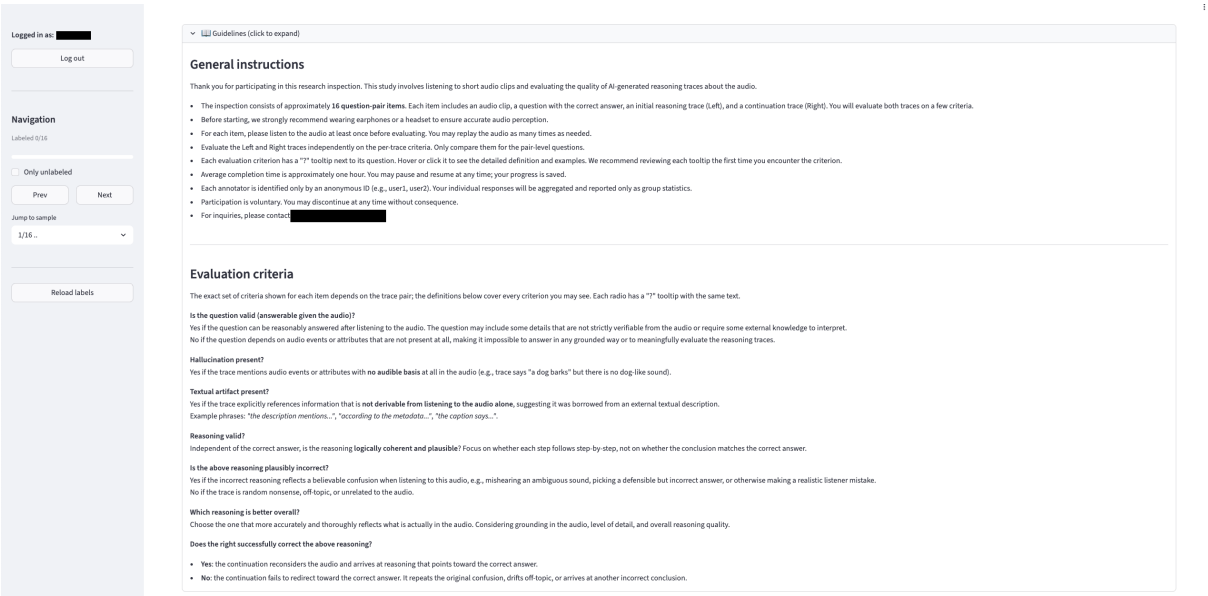


Figure 5: Full annotation guidelines shown to annotators at the start of the labeling interface.

Model	Overall	Perception	Cognition
Qwen2.5-Omni-7B	35.9	38.9	35.5
+ LLT (SFT → DPO)	**40.4**	**44.6**	**39.6**

Table 11: Results on WoW-Bench.

Model	MMAU Test-mini	MMAR
Qwen2.5-Omni-7B (reported)	71.5	56.7
Qwen2.5-Omni-7B (reproduced)	74.6	57.9
+ Gaussian noise	49.2	36.9
+ No audio	48.4	34.6
LLT (SFT → DPO)	78.2	63.7
+ Gaussian noise	48.9	31.4
+ No audio	58.7	39.3

Table 12: Audio ablation on MMAU Test-mini and MMAR.

replaced with a single judgment of *plausibly incorrect*: whether the incorrect reasoning reflects a believable confusion (e.g., mishearing an ambiguous sound) rather than random nonsense unrelated to the audio.

Pair-level evaluation. Annotators additionally provide a pair-level judgment depending on the type:

- **Type A: Better reasoning** (left / right) — Which reasoning is better overall, considering audio grounding, detail, and reasoning quality?
- **Type B: Successful correction** (Yes/No) — Does the continuation reconsider the audio and arrive at reasoning that points toward the correct answer?

	Type A		Type B	
	r^+	r^{++}	r^-	r^{++}
Per-trace quality				
Hallucination	13	19	–	18
Textual artifact	1	1	1	2
Reasoning valid	96	99	–	95
Plausibly incorrect	–	–	77	–
Pair-level judgment				
Better reasoning	35	**65**	–	–
Successful correction	–	–	**93**	

Table 13: Human inspection results (% of annotations on the valid-question subset, N=172). Question validity is 90% overall (Type A 88%, Type B 92%).

C.2 Results

The human inspection results are summarized in Table 13

Question validity. 90% of inspected questions

Item	% Agr.	Cohen's κ	Krippendorff's α
Question valid	85	0.22	0.22
Hallucination	79	0.21	0.22
Textual artifact	97	−0.01	−0.01
Reasoning valid	94	−0.03	−0.03
Plausibly incorrect	66	0.07	0.09
Better reasoning	50	−0.09	−0.08
Successful correction	88	−0.06	−0.05

Table 14: Inter-annotator agreement over the 96 doubly-annotated pairs.

were judged valid (88% for Type A, 92% for Type B).

Trace quality (Type A). Among valid pairs, annotators preferred the continuation r^{++} over the initial trace r^{+} on overall reasoning quality in 65% of cases (vs. 35% for r^{+}; forced choice, no tie option). Textual artifacts are rare ($\leq$2%) and reasoning is judged valid for nearly all traces (96% for r^{+}, 99% for r^{++}), indicating that continuation deepens reasoning while keeping it logically coherent.

Correction quality (Type B). 77% of the incorrect traces r^{-} were judged plausibly incorrect (realistic listener mistakes rather than random nonsense), supporting the design goal of Stage 3 for producing mistake-like errors. Notably, the continuation successfully corrected the reasoning in 93% of cases, providing direct human evidence that incorrect-to-correct traces ($r^{-} \rightarrow r^{++}$) reconsider the audio and arrive at the correct answer.

Hallucination. Annotation-level hallucination rates are 13% (r^{+}) to 18–19% (r^{++}), indicating that the continuation does not substantially increase hallucination despite producing longer and more detailed reasoning traces. Additionally, most flags come from a single annotator: when both annotators evaluate the same trace, they both flag hallucination in only 5% of cases, while at least one flags it in 26%. This suggests that the majority of flags reflect borderline judgments rather than clear-cut hallucination.

Inter-annotator agreement. We report raw agreement, Cohen's κ (Cohen, 1960), and Krippendorff's α (Krippendorff, 2011) in Table 14. Question validity and hallucination reach fair agreement ($\kappa \approx 0.22$, 79–85% raw agreement) under the Landis and Koch (1977) scale. Criteria with skewed label distributions (textual artifact, reasoning valid, successful correction) reach high raw agreement (88–97%) but near-zero κ/α due to the prevalence paradox. We rely on raw agreement for these items. The *better-reasoning* preference shows genuine disagreement (κ=−0.09, 50% raw agreement), reflecting the subjectivity of holistic quality judgments. We therefore treat the 65% Type-A preference as an aggregate tendency rather than a consensus.

D Examples from the LLT pipeline

D.1 Acoustic Description Example

The three examples below show model outputs when they are prompted to describe detailed acoustic characteristics of the given audio such as frequency, duration, amplitude, and timbre. The audio is from `ESC-50/3-151081-B-20` (baby crying), and the corresponding spectrogram is shown in Figure 7.

Qwen2-Audio-Instruct fails to follow the instruction and introduces unsupported acoustic claims. Gemini-2.5-Pro (Comanici et al., 2025) produces plausible-sounding but incorrect details, for example, asserting rapid modulation that is absent in the signal. In contrast, our Spectrogram+VLM approach yields the most accurate and grounded characterization among the three.

D.2 Stage 1 Description Example

The following figures show the three components of the multi-aspect dense description generated by Stage 1 for an example audio clip (VGGSound/-0jeONf82dE_000021): the semantic description (Figure 11), the acoustic description from spectrogram analysis (Figure 12), and the temporal description aggregating outputs from multiple specialized models (Figure 13).

D.3 Thought Expansion vs. Thought Continuation

The following reasoning traces correspond to the same example question, `YouTube8M/iBH70m_mjJw_143_10`, where `(D) Drums` is the correct answer. Although both approaches arrive at the correct prediction, Thought Expansion repeatedly restates the audio description and follows a more circuitous reasoning pattern. In contrast, Thought Continuation directly references auditory cues and produces a more focused and refined rationale.

E Prompts for LLT

This section provides the full prompts used at each stage of the LLT pipeline. For Stage 1, Figure 17 shows the prompt given to the vision-language

Source data	
Audio clips	338k
Original QA pairs (AF-Think)	338k
Hardened QA pairs (Stage 2)	316k
Total QA pairs	654k
Reasoning traces	
Short correct r^{+}	654k
Correct-to-refine $r^{+} \rightarrow r^{++}$	505k
Incorrect-to-correct $r^{-} \rightarrow r^{++}$	441k
Total	1.6M
Two-stage training	
SFT (r^{+}, $r^{+} \rightarrow r^{++}$)	1.16M
DPO preference pairs (20k $\times$ 3)	60k

Table 15: LLT dataset statistics. Counts are approximate.

model to produce the acoustic description from the spectrogram. For Stage 2, Figure 18 shows the prompt used to generate a more challenging question, and Figure 19 the prompt used to verify the generated question, reasoning, and answer. For Stage 3, Figure 20 shows the prompt used to synthesize an incorrect reasoning trace r^{-}. For Stage 4, Figure 21 shows the thought-continuation prompt, and Figure 22 the prompt used to verify and minimally repair the resulting long reasoning traces. All prompts are used with the open-source models described in Section 3.6.

F Training Details

For SFT, we perform full fine-tuning of all model components, including the LLM, the audio encoder, and the projection layers. We use a learning rate of 6×10^{-6}, batch size of 256, and train the model for 5 epochs. All SFT experiments are implemented using LLaMAFactory (Zheng et al., 2024).

Direct Preference Optimization (DPO) (Rafailov et al., 2023) learns the policy π_θ from pairwise preferences by increasing the likelihood of a preferred response o_w over a dispreferred response o_l, while regularizing against a fixed reference policy π_{ref} via a log-ratio contrast, as shown in Eq. 1.

We train DPO using a learning rate of 5×10^{-7} for 3 epochs with a batch size of 256. We use the sigmoid function $\sigma(\cdot)$ with a $\beta = 0.2$. We additionally include the SFT loss with a weight of 1.0 to stabilize training following Pang et al. (2024). All DPO experiments are implemented using Swift (Zhao et al., 2025b).

GRPO (Shao et al., 2024) uses the mean reward of responses sampled from the policy model as the baseline when computing the advantage. Specifically, a group of responses $\{o_1, o_2, \cdots, o_G\}$ is sampled from the policy model given the input question q, which yields corresponding rewards $\{r_1, r_2, \cdots, r_G\}$ from the reward model. The advantage is calculated over the mean:

$$\hat{A}_{i,t} = \widetilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}. \tag{2}$$

The policy model is further regularized with a Kullback–Leibler (KL) divergence penalty, as shown in Eq. 3.

Following common practice (Guo et al., 2025), we adopt a rule-based reward scheme that combines an accuracy reward for correct MCQ answers and a format reward for outputs that follow the `<think></think>` and `<answer></answer>` structure. The accuracy reward is assigned a weight of 1.0, while the format reward is weighted by 0.05.

We use a learning rate of 1.5×10^{-6}. For each prompt, we sample $G = 8$ responses from the policy model, set the batch size to 64, and train for 1,000 optimization steps. We use a sampling temperature of 1.0 and set the KL regularization coefficient β to 0.005. All GRPO experiments are implemented using Swift (Zhao et al., 2025b).

For every training, experiments are conducted on a single node with 8 NVIDIA H100 GPUs, and all reported results are from a single run.

G Evaluation Details

We mainly evaluate models on the widely used MMAU Test-mini (Sakshi et al., 2025), MMAR (Ma et al., 2025b), and MMSU (Wang et al., 2026a) benchmarks, which assess general audio understanding and reasoning capabilities across diverse tasks. MMAU evaluates information extraction and reasoning across speech, sound events, and music. MMAR focuses on complex, multi-step reasoning tasks spanning speech, sound, music, and their combinations. MMSU places greater emphasis on speech, evaluating both perceptual and reasoning abilities across paralinguistic and linguistic dimensions. For all benchmarks, we follow the official evaluation protocols: string matching is used for MMAU and MMAR, and letter matching is used for MMSU, with answers extracted from the `<answer></answer>` tags. To ensure fair comparison with previously reported results, we fix the benchmark version used for each: the `v05.15.25` release for MMAU Test-mini, and for MMAR and MMSU the annotations released before February 1, 2026, prior to subsequent updates to both benchmarks.

H Scientific Artifacts

Our work builds on open-source models and publicly available datasets, and we release the LLT dataset, the trained models, and the full data-construction pipeline. We enumerate the artifacts used and created below. To the best of our knowledge, our use of all artifacts is consistent with their intended research use, and we release our artifacts under terms compatible with those of the source data (research use only). License details are noted where available.

H.1 Models

- **Qwen2.5-Omni-7B** (Xu et al., 2025a)[2] — base LALM that we fine-tune (Apache-2.0).
- **Qwen3-Next-80B-Instruct**[3] / **Qwen3-Next-80B-Thinking**[4] — M_{instruct} / M_{reason} for data construction (Apache-2.0).
- **Qwen3-Omni-Captioner** (Xu et al., 2025b)[5] — semantic descriptions in Stage 1 (Apache-2.0).
- **Qwen3-VL-32B** (Bai et al., 2025)[6] — spectrogram interpretation in Stage 1 (Apache-2.0).
- **Qwen3-32B** (Yang et al., 2025a)[7] — text reasoner used only for the Stage-1 analysis (Apache-2.0).
- **Qwen2-Audio-Instruct** (Chu et al., 2024)[8] / **Qwen3-Omni-Instruct** (Xu et al., 2025b)[9] — additional backbones for the generalization study (Apache-2.0).
- **faster-whisper** (Radford et al., 2023)[10] (MIT), **pyannote.audio** (Bredin, 2023)[11] (MIT), and a **BEATs**-based sound event detector (Chen et al., 2023; Schmid et al., 2025)[12] (MIT) — temporal information in Stage 1.
- **allin1** (Kim and Nam, 2023)[13] (MIT) and **autochord** (Bayron, 2023)[14] (MIT) — music structure and chord recognition in Stage 1.
- **LLaMA-Factory** (Zheng et al., 2024)[15] (Apache-2.0) for SFT and **ms-swift** (Zhao et al., 2025b)[16] (Apache-2.0) for DPO/GRPO; **LoRA** (Hu et al., 2022) for the generalization study.
- **GPT-4.1-mini**[17] — cognitive behavior classification and correction analysis in Section 4.2 (proprietary; OpenAI Terms of Use).

H.2 Datasets

- **AF-Think** (Goel et al., 2025)[18] — the source audio CoT dataset we enrich (NVIDIA OneWay Noncommercial License). For reproducibility we exclude licensed subsets such as **Music4all** (Santana et al., 2020) and limit audio to 30 seconds. The underlying audio originates from public datasets (e.g., LibriSpeech, VGGSound, AudioSet, Clotho, DailyTalk, MELD, FMA, ESC-50).
- **AVQA** (Yang et al., 2022)[19] — audio question-answering data for the GRPO stage.
- **AudioMCQ** (He et al., 2025)[20] — alternative source dataset for the generalization study (Apache-2.0).
- **MMAU** (Sakshi et al., 2025), **MMAR** (Ma et al., 2025b), **MMSU** (Wang et al., 2026a), **WoW-Bench** (Kim et al., 2025), and **TREA** (Bhattacharya et al., 2025) — evaluation benchmarks.

[2] https://github.com/QwenLM/Qwen2.5-Omni
[3] https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct
[4] https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Thinking
[5] https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Captioner
[6] https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct
[7] https://huggingface.co/Qwen/Qwen3-32B
[8] https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct
[9] https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct
[10] https://github.com/SYSTRAN/faster-whisper
[11] https://github.com/pyannote/pyannote-audio
[12] https://github.com/fschmid56/PretrainedSED
[13] https://github.com/mir-aidj/all-in-one
[14] https://github.com/cjbayron/autochord
[15] https://github.com/hiyouga/LLaMA-Factory
[16] https://github.com/modelscope/ms-swift
[17] https://platform.openai.com/docs/models/gpt-4.1-mini
[18] https://huggingface.co/datasets/nvidia/AF-Think
[19] https://github.com/AlyssaYoung/AVQA
[20] https://huggingface.co/datasets/inclusionAI/AudioMCQ

H.3 Released Artifacts

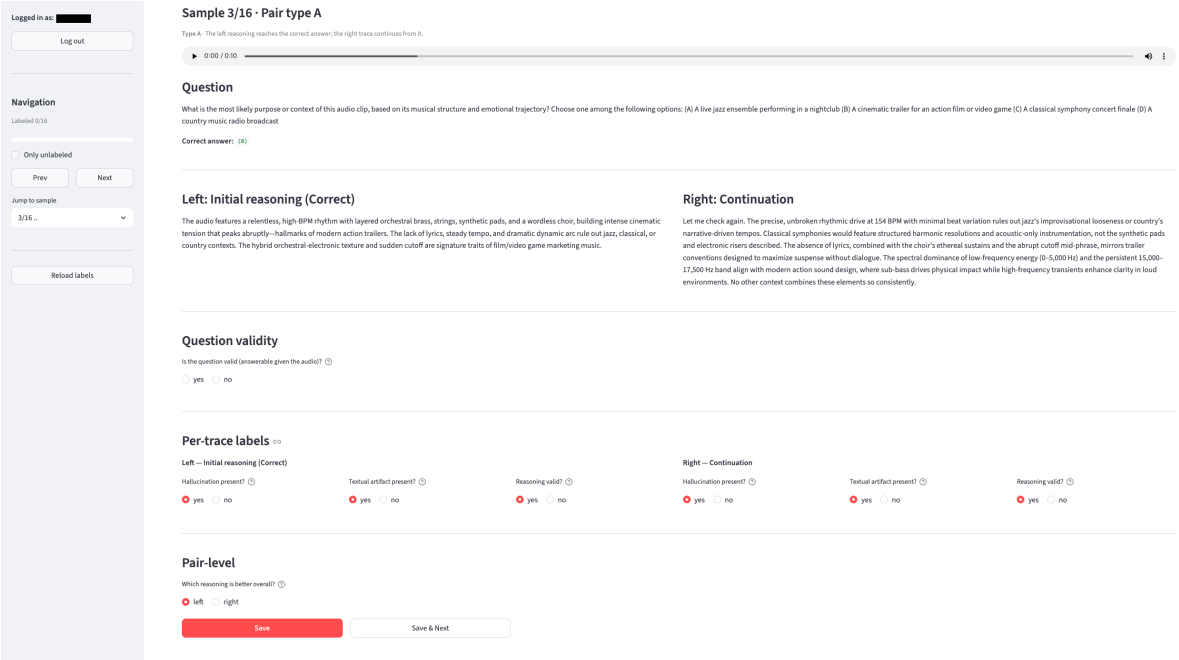
We release (i) the **LLT dataset** of approximately 1.6M reasoning traces (statistics in Table 15), (ii) the **LLT-trained models**, and (iii) the full **data-construction pipeline**. These are released for research use, consistent with the research-use terms of the source datasets; derivatives of research-only data are not intended for use outside research contexts.

H.4 Data Statistics, Coverage, and Privacy

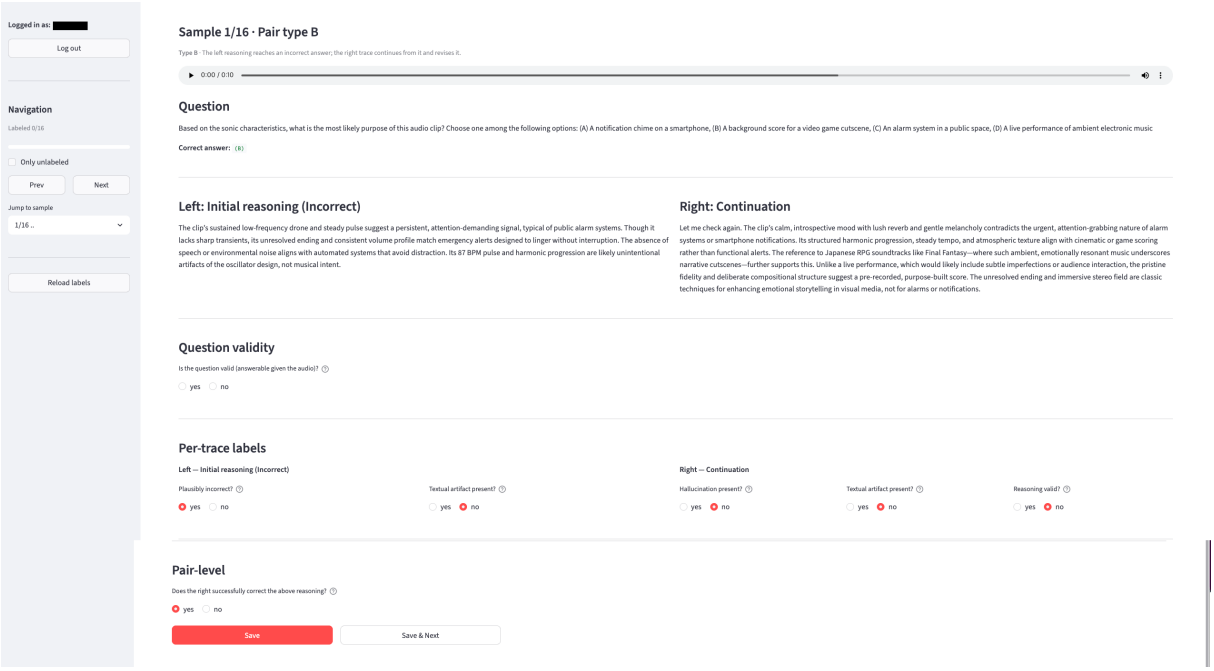
The LLT dataset spans speech, sound events, and music; per-component statistics and training splits are reported in Table 15 and Appendix F. All audio and question–answer pairs are drawn from existing public research datasets; we do not collect new data or add personally identifying metadata. Speech recordings in these datasets may contain individual voices, but we introduce no new identifying annotations. The human inspection (Appendix C) was conducted by annotators identified only by anonymous IDs, with all responses aggregated into group statistics.

I Use of AI Assistants

We used AI assistants (large language models) only for minor coding support and language editing. All research ideas, experiments, analyses, and claims are the authors' own, and the authors reviewed and verified all AI-assisted code and text.



(a) Type A pair $(r^{+},\ r^{+} \rightarrow r^{++})$.



(b) Type B pair $(r^{-},\ r^{-} \rightarrow r^{++})$.

Figure 6: Labeling interface presented to annotators for (a) Type A and (b) Type B trace pairs. Each item shows the audio clip, the question with its correct answer, the initial trace (left) and the continuation (right), and the per-trace and pair-level judgments.

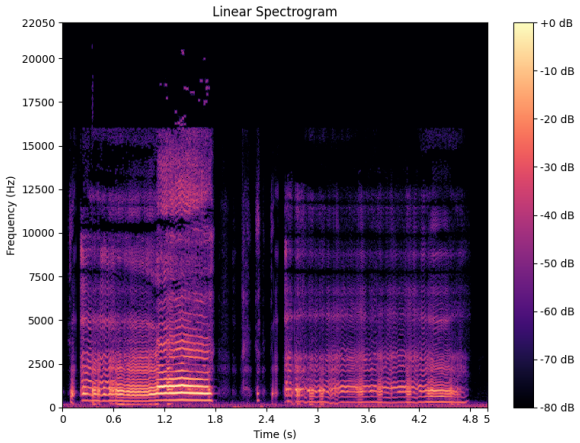


Figure 7: Spectrogram of `ESC-50/3-151081-B-20` (Baby Crying)

Description from Qwen2-Audio-Instruct

The audio features a male singing in English with a sad mood. The characteristics include a key of Bb minor, time signature of 4/4, and a tempo of 150.0 bpm.

Figure 8: Acoustic description generated by Qwen2-Audio-Instruct.

Description from Gemini-2.5-Pro

The audio is a 4-second snapshot of a complex vocal event. While short, it contains a significant amount of acoustic information that our brains immediately interpret as a specific emotional state: distress. Here is a more detailed look at the components:

Frequency (Pitch)

Frequency corresponds to our perception of pitch. In this audio, the cry is characterized by:

High Fundamental Frequency: The baseline pitch is high, which is a direct result of the small and thin vocal folds of an infant. Just as a thin guitar string produces a higher note than a thick one, an infant's smaller vocal anatomy produces a higher-pitched sound.

Wide and Rapid Fluctuation (Pitch Contour): The pitch is not steady. It sweeps up and down rapidly, creating the classic "wailing" sound. This variation, or pitch contour, is a key acoustic cue for emotion. The initial sound starts at a certain pitch, then rises sharply, and falls again in a pattern that conveys urgency and discomfort. This dynamic shifting is far more attention-grabbing than a monotone sound would be...

Figure 9: Acoustic description generated by Gemini-2.5-Pro.

Description from Spectrogram + Qwen3-VL-32B

Duration: The audio spans approximately 4.5 seconds, as indicated by the horizontal axis (time scale from 0 to 4.5). The signal is continuous with distinct segments: a sustained sound from 0.0 to 1.8 seconds, followed by a brief silence or low-energy period (1.8–2.3 seconds), and then a return of similar sustained sound from 2.3 to 4.5 seconds.

Frequency: The frequency range extends from 0 Hz to approximately 19,845 Hz, as shown on the vertical axis. The dominant energy is concentrated in the lower to mid-frequency range (below 11,025 Hz), particularly between 2,205 Hz and 6,615 Hz, where dense, periodic horizontal striations are visible – indicative of harmonic structure. There are also consistent, weaker harmonics extending into the higher frequencies (up to 15,435 Hz), especially during the first segment. **A few transient high-frequency components appear around 1.0–1.5 seconds at 17,640–19,845 Hz, suggesting possible percussive or formant-like events.** Between 1.8–2.3 seconds, there is a noticeable drop in spectral content, indicating a pause or reduction in vocal/sonic activity...

Figure 10: Acoustic description generated by our Spectrogram + Qwen3-VL-32B approach.

Stage 1 Example: Semantic Description (VGGSound/-0jeONf82dE_000021)

The audio opens with the unmistakable sound of a powerful V8 engine revving, its deep, throaty roar filling the space and immediately conjuring a sense of American muscle and automotive culture. Over this mechanical backdrop, a male voice—gruff, gravelly, and distinctly Southern—asks, “Do you smell barbecue?” His delivery is casual and relaxed, evoking the atmosphere of a Southern cookout or gathering. As the engine sound fades, a second male voice, bright and energetic, exclaims “Hulk!” with a playful, comic-book inflection, referencing the iconic Marvel character and introducing a whimsical, pop-culture element to the scene.

The tone shifts as a third voice, calm and smooth, delivers the tagline “Fresh, smooth, real. Bud Light.” The voice is polished and resonant, set against a clean, studio-produced instrumental bed featuring a funky, syncopated bass line, crisp drum machine beat, and subtle electric guitar or keyboard chords. The music and voice are tightly mixed, with the jingle’s rhythm and melody reinforcing the brand’s identity and creating a memorable, catchy advertisement. This segment is unmistakably a commercial for Bud Light beer, crafted to evoke a sense of coolness and authenticity.

Suddenly, the mood pivots again as a fourth voice, high-pitched and enthusiastic, exclaims, “Cooler rocketslide!” The delivery is reminiscent of a young boy or a character from a children’s animated show, introducing a playful, imaginative twist. The music continues, but the tone is now more lighthearted and comedic, suggesting a family-friendly or youth-oriented context. The final voice, deeper and more formal, calmly states, “It’s all here,” serving as a closing statement that ties the varied elements together and signals the end of the commercial.

The audio clip is a tightly produced, high-fidelity advertisement that blends automotive, barbecue, comic-book, and pop-culture references with the recognizable Bud Light brand jingle. It features a sequence of distinct male voices—each with unique regional and stylistic traits—delivered in a montage that moves from Southern barbecue to Marvel fandom, from beer commercial to playful children’s exclamation. The music shifts from raw rock to funky pop, maintaining crisp, professional sound quality throughout.

Figure 11: Stage 1 example: semantic description generated by Qwen3-Omni-Captioner.

Stage 1 Example: Acoustic Description (VGGSound/-0jeONf82dE_000021)

Overview: The spectrogram displays a linear frequency scale from 0 to 22,050 Hz and a time span of 0 to 9.01 seconds. The intensity is represented by a color gradient ranging from black (-80 dB) to yellow (0 dB).

Frequency: Energy is concentrated primarily between 0 and 15,000 Hz, with the most intense components occurring between 2,500 and 12,500 Hz. Below 2,500 Hz, energy is present but less intense. Above 15,000 Hz, activity is sparse and appears as intermittent, low-intensity vertical streaks.

Amplitude/Intensity: The highest intensities are observed in the mid-frequency range (approximately 2,500–12,500 Hz) during periodic time intervals, reaching levels around -10 to 0 dB. Between these peaks, the signal drops to around -40 to -60 dB, with background noise near -70 to -80 dB.

Timbre: The signal exhibits a complex, broadband timbre with a strong presence of harmonic-like structures, particularly evident in the form of vertical striations and clustered bands.

Duration of events: Acoustic activity occurs in recurring, short-duration bursts that last approximately 0.5 to 1.0 seconds each, separated by brief gaps of lower intensity, appearing roughly every 1.0 to 1.5 seconds.

Onset/Offset Characteristics: Each burst exhibits a rapid onset, typically within 0.1–0.2 seconds, marked by a sudden increase in intensity across a broad frequency band. Offsets are similarly abrupt.

Spectral Characteristics: The spectrum features distinct vertical striations and horizontal banding in the 2,500–12,500 Hz range, indicating a quasi-periodic structure with possible harmonic content.

Spatial/Reverberation: No clear reverberant tails or echo patterns are visible, indicating minimal room reverberation or a direct, close-field recording condition.

Periodicity: There is a clear periodicity in the temporal domain, with bursts occurring at approximately 1.0–1.5 second intervals throughout the recording.

Noise Components: Background noise is relatively low, with large portions of the spectrogram near the -70 to -80 dB level.

Summary: The signal consists of a series of short, high-intensity, broadband bursts occurring at regular intervals over 9.01 seconds, with a complex, harmonically rich spectral structure concentrated between 2,500 and 12,500 Hz, suggesting a modulated or pulsed acoustic source with minimal reverberation.

Figure 12: Stage 1 example: acoustic description generated by Qwen3-VL-32B from the spectrogram.

Stage 1 Example: Temporal Description (VGGSound/-0jeONf82dE_000021)

Sound Event Detection (Threshold 0.2):
[0.00–0.84 s] Neigh, whinny
[0.76–1.40 s] Background noise
[0.80–2.96 s] Male speech, man speaking
[1.24–2.40 s] Wind
[2.08–2.32 s] Background noise
[2.76–3.00 s] Wind
[3.04–9.01 s] Music
[3.36–7.44 s] Male speech, man speaking
[7.28–8.68 s] Neigh, whinny
[8.24–9.01 s] Male speech, man speaking

Sound Event Detection (Threshold 0.5):
[0.20–0.52 s] Neigh, whinny
[0.88–2.80 s] Male speech, man speaking
[3.12–9.00 s] Music
[3.44–5.88 s] Male speech, man speaking
[6.16–7.32 s] Male speech, man speaking
[7.40–7.96 s] Neigh, whinny
[8.32–9.01 s] Male speech, man speaking

Speech Transcription with Timestamps:
[0.50–1.86 s] “Do you smell barbecue?”
[2.34–2.82 s] “Oh!”
[3.34–5.86 s] “Fresh, smooth, real, Bud Light.”
[6.14–7.38 s] “Cool a rocket sleigh.”
[8.22–9.00 s] “It’s all here.”

Speaker Diarization:
[0.44–3.12 s] SPEAKER_00
[0.84–1.01 s] SPEAKER_01
[3.41–9.09 s] SPEAKER_00

Music Structure Analysis:
Tempo (BPM): 128.0
Beats Count: 19, Downbeats Count: 5
Beat Start Time: 0.75 s, Downbeat Start Time: 0.75 s
Average Beat Interval: 0.461 s, Std Dev: 0.028 s
Beat Stability: steady
Music Sections: [0.0–10.0 s] verse

Chord Recognition:
[2.93–5.94 s] A:min

Figure 13: Stage 1 example: temporal description aggregating timestamps from Whisper, pyannote, BEATs-based sound event detection, allin1 music structure analysis, and autochord chord recognition.

Example Question and Initial Reasoning Trace

Question
Which instrument provides the rhythmic foundation in the audio? Choose one among the following options:
(A) Harmonica
(B) Flute
(C) Violin
(D) Drums
Please think and reason about the input audio before you respond.

Initial Incorrect CoT Trace
The audio features a persistent, pulsing low-frequency tone that resembles a bowed string instrument's sustain, and the rhythmic chant aligns closely with the vibrato-like modulation in the bassline. While drums are present, they're masked by heavy compression. The violin's melodic contour, though subtle, drives the groove with syncopated bowing patterns, making it the true rhythmic anchor.

Figure 14: Example question and initial incorrect CoT trace used to compare thought expansion and thought continuation. Correct answer: (D) Drums.

Example of Thought Expansion

Let me listen again. The spectrogram shows strong energy between 0 and 15 kHz with no harmonic peaks. The timbre is broadband and noise-like. The temporal description shows music and male singing throughout, but no specific mention of a violin. The acoustic description says no harmonic structure or periodicity. **The semantic description says** "powerful, four-on-the-floor kick drum, sharp synthetic snare hits, and a pulsating, low-frequency bassline that forms the core of the groove." The question is about the rhythmic foundation. The rhythmic foundation is typically the kick drum in dance music. The bassline supports it, but the kick drum provides the steady pulse. Let me check the options: Harmonica, Flute, Violin, Drums. **The semantic description clearly states** "powerful, four-on-the-floor kick drum, sharp synthetic snare hits, and a pulsating, low-frequency bassline that forms the core of the groove." So the kick drum is part of the drums. The bassline is also part of the rhythmic foundation, but the drums (kick and snare) are the primary rhythmic elements. The bassline is more of a melodic and harmonic element, but it also has rhythmic aspects. However, in dance music, the kick drum is the main rhythmic foundation. **The acoustic description says:** "Energy is distributed across the entire frequency range, but the most intense components are concentrated between 0 and 15,000 Hz." The low-frequency energy is from the kick and bass. **The temporal description shows** music and male singing, but not specifically drums. **But the semantic description explicitly says** "powerful, four-on-the-floor kick drum, sharp synthetic snare hits" which are drums. The question is "which instrument provides the rhythmic foundation". In electronic dance music, the rhythmic foundation is typically the kick drum. The options are Harmonica, Flute, Violin, Drums. Drums would include the kick and snare. The bassline is not an instrument in the options. So Drums would be the correct choice. The other options are all melodic instruments: harmonica, flute, violin. They are not typically used for rhythmic foundation in this context. The acoustic description says "no harmonic structure" which would suggest it's not a melodic instrument. Drums are percussive and non-harmonic. So the rhythmic foundation must be drums. So the answer is (D) Drums.

<answer> (D) </answer>

Figure 15: Example of thought expansion: M_{reason} repeatedly restates the audio description and follows a circuitous reasoning pattern.

Example of Thought Continuation

Hmm. The audio features a consistent, driving beat with a prominent kick drum pattern that pulses on every beat, forming the backbone of the rhythm. The high-energy electronic percussion includes sharp snare hits and steady hi-hats, but the most dominant rhythmic element is the deep, repetitive kick drum that anchors the entire track. There are no melodic or harmonic instruments like harmonica, flute, or violin present—the sound is purely synthetic and percussive, with no acoustic string or wind instruments detectable. The bassline, while low-frequency, is part of the electronic production and not a violin or other traditional instrument. **The drums, specifically the kick drum, are unmistakably the rhythmic foundation**.

<answer> (D) </answer>

Figure 16: Example of thought continuation: M_{reason} directly references auditory cues and produces a more focused, refined rationale.

Prompt for Acoustic Description (Stage 1)

System Prompt:
You are an expert in acoustics and audio signal analysis. You analyze spectrograms with high precision based on the visual evidence shown in the image. Your descriptions must be concise, technically accurate, and fully aligned with the user's instructions. You must follow the required output format exactly.

User Prompt:
Please provide an acoustic description of the given spectrogram. Describe key features such as frequency, amplitude/intensity, timbre, and duration of events. When relevant, you may also comment on additional characteristics such as reverberation, periodicity, spectral features, onset/offset characteristics, harmonic structure, and noise components. When using numeric values, ensure they are reported accurately according to the axis labels and grid lines. Do not infer or suggest what the audio event represents.
If a particular characteristic is not clearly observable in the spectrogram, you can leave the corresponding field as an empty string ("").
Return your response in the following JSON format, following the structure and level of detail illustrated in the example below.

```
{
"Overview": "", "Frequency": "", "Amplitude/Intensity": "", "Timbre": "",
"Duration of events": "", "Onset/Offset Characteristics": "",
"Spectral Characteristics": "", "Spatial/Reverberation": "",
"Periodicity": "", "Noise Components": "", "Others": "", "Summary": ""
}
```

Example:

```
{
"Overview": "The spectrogram displays a linear frequency scale from 0 to 24,000 Hz and a time span of 0 to 18 seconds. The intensity is represented by a color gradient ranging from black (−80 dB) to yellow (0 dB).",
"Frequency": "Prominent energy is concentrated below 10,000 Hz, with the most intense components occurring between 0 and 5,000 Hz.",
"Amplitude/Intensity": "The highest intensities are localized primarily in the lower-frequency range below 3,000 Hz and occur in discrete time intervals.",
"Timbre": "Complex, predominantly broadband spectral texture with irregular energy distributions.",
"Duration of events": "Approximately 18 seconds, with intermittent short bursts of activity.",
"Onset/Offset Characteristics": "Events exhibit rapid onsets and offsets with sharp transitions.",
"Spectral Characteristics": "Dominated by broadband energy with irregular spectral shapes.",
"Spatial/Reverberation": "No clear reverberant tails or echo patterns are visible.",
"Periodicity": "No stable periodic structure is observed.",
"Noise Components": "Background noise remains near the noise floor.",
"Others": "", "Summary": "Short-duration broadband pulses dominate the signal."
}
```

Figure 17: Stage 1 prompt used with Qwen3-VL-32B to generate the acoustic description from the spectrogram image.

Prompt for Question Generation (Stage 2)

System Prompt:
You are a question generation assistant. Your task is to generate more challenging questions based on audio and original questions. You must strictly follow the user's instructions and output format. Return ONLY valid JSON and nothing else.

User Prompt:
You are tasked with generating a MORE CHALLENGING question based on the given audio description and the original question.
<original_question> __ORIGINAL_QUESTION__ </original_question>
<original_response> __ORIGINAL_RESPONSE__ </original_response>
<audio_description> __AUDIO_DESCRIPTION__ </audio_description>

Guidelines:

- The respondent will be given only the audio and the question. Do NOT explicitly mention or refer to the existence of any description or metadata. The reasoning should sound as if it comes solely from listening to the audio.
- Preserve the same question format as the original whenever possible (multiple-choice or free-form).
- Increase difficulty along one or both of the following dimensions:
  - More audio-aware: requires closer, more precise listening.
  - More complex reasoning: requires additional reasoning steps (e.g., combining cues, resolving ambiguity, eliminating plausible alternatives).
- Prefer to retain the core skill(s) tested by the original question. If preserving the same skill would force an unnatural question, design a new well-grounded question based on the same audio.
- The question must be answerable from the audio alone.
- For multiple-choice: create plausible distractors similar in difficulty and surface plausibility, while ensuring the correct answer can still be determined through careful listening.
- Along with the question, generate (i) a short chain-of-thought (max 100 words) leading to the correct answer, and (ii) the final answer.

Output: Return your output strictly in the following JSON format:
{
"question": "<generated question>",
"response": "<short reasoning trace leading to the answer>",
"answer": "<choice letter for MCQ, or concise text for free-form>",
"choices": null or ["choice A", "choice B", "choice C", "choice D", ...]
}

Figure 18: Stage 2 prompt used with M_{instruct} to generate a more challenging question from the original QA pair and the dense audio description.

Prompt for Question Verification (Stage 2)

System Prompt:
You are a verification assistant. Your role is to evaluate the provided generated question according to the given verification criteria and produce a JSON judgment in the specified output format.

User Prompt:
You are a verifier for audio-based questions. You will be given (i) an audio description (used only as a proxy for the audio), (ii) a question generated based on the audio, (iii) a model-generated reasoning trace, and (iv) a final answer.

Your task is to verify whether the generated question and response satisfy the verification criteria below. Treat the audio description strictly as a proxy for the audio. Your judgment should reflect whether the question can be reasonably answered based on the given audio.

Verification Criteria

1. Audio-based Solvability. Determine whether the question is grounded in the audio signal and can be answered using audible evidence. A question is NOT audio solvable if:

- The required distinction is ambiguous or indistinguishable from the audio signal alone.
- The question relies on hallucinated events/attributes that cannot be identified by listening.
- The question relies on over-interpretation of the audio description.
- The question depends on attributes not defined by the audio (e.g., visual details, metadata).
- For multiple-choice: distractors cannot be reliably distinguished, or more than one choice remains equally plausible.

2. Answer Correctness. For multiple-choice, the selected answer must be correct relative to the provided choices. For non-multiple-choice, the free-form answer must be correct.
3. Reasoning Validity. The reasoning must be logically coherent, plausibly justify the final answer based on audible evidence, and must not mention or rely on the audio description or metadata. Minor stylistic issues are acceptable.

Output Format: Return your judgment strictly in the following JSON format:

```
{
"audio_based_solvable": true or false,
"answer_correct": true or false,
"reasoning_valid": true or false
}
```

Input

```
<audio_description> __AUDIO_DESCRIPTION__ </audio_description>
<question> __QUESTION__ </question>
<reasoning> __REASONING__ </reasoning>
<answer> __ANSWER__ </answer>
```

Figure 19: Stage 2 prompt used with M_{instruct} to verify the generated question, reasoning, and answer.

Prompt for Incorrect CoT Generation (Stage 3)

System Prompt:
You are an expert reasoning assistant. Your job is to produce coherent, logically structured reasoning traces according to the user's instructions. You must follow the required output format precisely.

User Prompt:
<question> __QUESTION__ </question>
<original_cot> __ORIGINAL_COT__ </original_cot>
<audio_description> __AUDIO_DESCRIPTION__ </audio_description>

Above is a question and its original chain-of-thought (CoT) response that leads to the correct answer. Additional audio descriptions are also provided above. You may refer to these descriptions if needed when generating the reasoning trace.

Your task is to generate an alternative reasoning trace that sounds logical and plausible but leads to an incorrect final answer.

Requirements for the incorrect reasoning trace:

- It must be coherent, step-by-step, and internally consistent.
- It should appear reasonable on the surface but include subtle mistakes, incorrect assumptions, misinterpretations of the audio (e.g., misperceiving timing or events), or overlooked details.
- It must NOT contradict facts randomly; the reasoning should feel believable.
- The incorrect answer must be a valid option for the question if it is multiple-choice.
- Do NOT copy the original chain-of-thought. Produce a new and distinct reasoning process.
- Do NOT change the wording of the question or the given choices.
- The length should be similar to the original chain-of-thought and must not exceed ~100 words.

Answer-format requirements: For multiple-choice with explicit letters, use ONLY the choice letter. For multiple-choice without letters, use the exact choice text. For free-form, provide a concise free-form answer.

Return your output strictly in the following JSON format:
{
"question": "<insert question>",
"original_answer": "<insert original correct answer>",
"wrong_answer": "<insert incorrect final answer>",
"wrong_response": "<insert step-by-step reasoning leading to the incorrect answer>"
}

Figure 20: Stage 3 prompt used with M_{instruct} to synthesize an incorrect reasoning trace r^- from the original (q, r^+, a^+) and the dense audio description.

Prompt for Thought Continuation (Stage 4)

System Prompt:
You are an assistant whose role is to improve and correct reasoning.

User Prompt:
<audio_description> __AUDIO_DESCRIPTION__ </audio_description>

Answer the question below as if you are listening directly to the audio depicted by the description above. You may refer to that description for accuracy, but you must not explicitly mention or quote it. Avoid statements like "The description mentions. . . " inside your response.

You will be given a previously generated chain-of-thought inside a <reasoning> block. Continue the reasoning to solve the question accurately. You may correct any errors, incorrect assumptions, or flaws in the reasoning if they exist. Continue exactly from the last word of the given chain-of-thought, as if you are completing the unfinished sentence.

Your continuation should:

- follow naturally from the text already written.
- always extend the reasoning, even if you believe the original final answer is already correct.
- use the following cognitive behaviors when helpful: verification (systematically checking intermediate results), backtracking (explicitly revising approaches when errors are detected), subgoal setting (breaking a complex problem into manageable steps), and backward chaining (reasoning backward from a desired outcome to the initial inputs).
- optionally revise the final answer based on your improved reasoning. The answer may remain the same or change.
- NOT mention or refer to the description explicitly; the reasoning must sound as if it comes solely from listening to the audio.

Place the final answer inside <answer>. . . </answer> after the reasoning.

Answer format: For multiple-choice with explicit letters, use ONLY the letter, e.g., <answer> (B) </answer>. For multiple-choice without letters, use the exact choice text, e.g., <answer> Grief </answer>. For non-multiple-choice, use a concise answer, e.g., <answer> To achieve political goals by violent means </answer>.

<question> __QUESTION__ </question>

Figure 21: Stage 4 prompt used with M_{reason} to continue an initial reasoning trace $r \oplus c$ with explicit cognitive behaviors.

Prompt for Long Reasoning Trace Verification (Stage 4)

User Prompt:
You are a verifier for reasoning traces on audio-based questions. You will be given (i) an audio description (used only as a proxy for the audio), (ii) a question, (iii) the correct answer, (iv) a previous reasoning (context), and (v) a continued reasoning generated by the model that is intended to follow the previous reasoning.

IMPORTANT: Evaluate ONLY the <continued_reasoning>. The <previous_reasoning> is provided only as context and may be correct or incorrect; do NOT judge or correct it. Your role is strictly to verify, not to rewrite or improve the content.

Verification Criteria

1. Answer Correctness. The final answer in the continued reasoning must match the correct answer. For free-form, preserve the core meaning (differences in specificity or phrasing are acceptable; do NOT rewrite the <answer> to match the correct answer verbatim). For multiple-choice with explicit letters, the choice letter must be correct; without letters, the selected text must match the correct choice.
2. Reasoning Validity. The continued reasoning must be logically coherent and consistent and plausibly support the final answer. Clear logical contradictions, hallucinated claims, or unsupported reasoning steps invalidate the reasoning.
3. Format Validity.

- Description / Metadata Mentions: Must NOT explicitly mention the description or metadata (e.g., "the description says. . . ", "according to the description. . . "). The reasoning must sound as if it comes solely from listening to the audio.
- Answer Tag Rules: The final answer must appear inside the final <answer>. . . </answer> tag, and the tag must appear only once.
- Answer Format Rules: For free-form, preserve the original answer if correct. For MCQ with letters: <answer> (letter) </answer>. Without letters: <answer> choice text </answer>. For non-MCQ: <answer> concise answer </answer>.
- Forbidden Tags: No other tags (e.g., <reason>, <think>, <reasoning>) should appear in the continued reasoning.

Format Validation Policy: If violations are minor and fixable (incorrect answer formatting, removable extra tags, or rewritable description mentions), set "format_valid": true and place the minimally corrected version in "fixed". If violations are severe (multiple conflicting answers, unremovable dependence on the description), set "format_valid": false and "fixed": null.

Figure 22: Stage 4 prompt used with M_{instruct} to verify and (where possible) minimally repair long reasoning traces generated via thought continuation.

$$\mathcal{J}_{\text{DPO}}(\theta) = \mathbb{E}_{(q,o_w,o_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(o_w \mid q)}{\pi_{\text{ref}}(o_w \mid q)} - \beta\log\frac{\pi_\theta(o_l \mid q)}{\pi_{\text{ref}}(o_l \mid q)}\right)\right]. \tag{1}$$

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Bigg(\min\Big[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}\hat{A}_{i,t}, \\ \text{clip}\Big(\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, 1-\epsilon, 1+\epsilon\Big)\hat{A}_{i,t}\Big] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta \,\|\, \pi_{\text{ref}}]\Bigg)\Bigg]\end{aligned} \tag{3}$$