

Perception Is the Bottleneck for Large Audio Reasoning Models

Ziyang Ma^{1,2,3}, Ruiyang Xu¹, Wenming Tu¹, Zhenghao Xing⁴, Yuxuan Wang⁵,
Yunfei Chu⁵, Jin Xu⁵, Kai Yu¹, Junyang Lin⁵, Eng-Siong Chng², Xie Chen^{1,3†}

¹ X-LANCE Lab, Shanghai Jiao Tong University, ² Nanyang Technological University,

³ Shanghai Innovation Institute, ⁴ The Chinese University of Hong Kong

⁵ Qwen Team, Alibaba Group

Abstract

While Chain-of-Thought (CoT) reasoning has revolutionized text-based LLMs, its application to the audio modality reveals a counter-intuitive thinking paradox: forcing models to “think” via prompting or post-training often degrades performance compared to direct answering. In this work, we investigate the bottleneck of audio reasoning and reveal a critical insight: the primary constraint is not the model’s logical deduction capability, but its inability to accurately perceive acoustic details. We demonstrate that keeping perceptual cues effectively activates the LALM’s internal reasoning potential. Building on this finding, we propose **Audio-ReaPer**, a post-training framework and corresponding model designed to unlock Reasoning potential by explicitly fortifying Perception. Our approach reforms the post-training paradigm through three key strategies: (1) utilizing agentic data generation pipeline to synthesize perception-enhanced data for SFT; (2) curating perception-focused data for RLVR; and (3) introducing a novel perceptual density reward during RL to stabilize the learning process and balance observation with deduction. Extensive experiments show that Audio-ReaPer establishes a new state-of-the-art on benchmarks such as MMAU, MMAU-Pro, and MMAR. Notably, our 7B model outperforms previous open-source models by significant margins and surpasses proprietary models like GPT-4o Audio and Gemini 2.5 Flash. The model and code will be open-source to facilitate further research for audio reasoning.

1 Introduction

Multimodal reasoning serves as a cornerstone for AI systems to solve complex real-world problems. With the rapid evolution of Large Audio Language Models (LALMs) (Chu et al., 2023; Tang et al., 2024; Kong et al., 2024), capabilities in basic auditory tasks such as speech recognition and

sound event detection have been achieved. The field is now pivoting towards a more ambitious goal: training Large Audio Reasoning Models (LARMs) with complex reasoning capabilities to tackle complicated acoustic scenarios, such as analyzing mixtures of speech, music, and environmental sounds (Ma et al., 2025b), where simple pattern matching falls short.

Significant research efforts have been dedicated to this pursuit, ranging from in-context prompting (Ma et al., 2025a) to supervised fine-tuning (Xie et al., 2025) with Chain-of-Thought (CoT) data. Inspired by the breakthrough of reasoning models like DeepSeek-R1 (Guo et al., 2025) in the text domain, recent works have begun applying Reinforcement Learning with Verifiable Rewards (RLVR) to the audio modality (Li et al., 2025; Rouditchenko et al., 2025). The prevailing assumption is that by scaling CoT data or applying Group Relative Policy Optimization (GRPO) directly to audio question answering, reasoning capabilities will spontaneously emerge in LALMs. However, empirical observations contradict this expectation. Recent studies (Li et al., 2025) reveal a counter-intuitive phenomenon: models forced to employ a “thinking” process during RL post-training often underperform their direct-answering counterparts. While some work attempt to mitigate this via iterative distillation (Tian et al., 2025) or adaptive computation (Wu et al., 2025), they often treat the symptom rather than the root cause. A fundamental question remains unanswered:

What exactly is the bottleneck for audio reasoning?

In this work, we address this question through a rigorous bottleneck analysis comprising error pattern analysis and oracle injection study. Our error analysis reveals that even for state-of-the-art

[†]Corresponding author

LALM, the most abundant part of reasoning failures stems from perception errors rather than logical fallacies. Furthermore, our oracle injection study uncovers a pivotal finding: when a model is forced to think but provided with ground-truth perceptual cues, its performance improves dramatically. This leads to a critical insight: **the capacity for audio reasoning is largely intrinsic to the LLM backbone; the true bottleneck lies in the LARM’s inability to establish perception-reasoning alignment, in the lack of precise evidence upon which to reason.**

Driven by this insight, we propose **Audio-ReaPer**, a novel framework that fundamentally reforms the post-training paradigm for audio Reasoning via enhancing **Perception**. We hypothesize that a robust reasoning chain cannot be forged in isolation but must be built upon a foundation of high-fidelity perception. To realize this, we introduce three synergistic contributions:

1. **Perception-Enhanced SFT:** We depart from the standard cold-start method, which often suffers from sparse supervision. Instead, we utilize an agentic data generation pipeline to synthesize detailed, high-density descriptive data. This initializes the model with a “strong ear,” ensuring it can exhaustively describe acoustic events before attempting to solve complex tasks.
2. **Perception-Focused RLVR:** We construct a specialized RL environment designed to trigger audio reasoning. By generating multi-cue dependency questions that require distinct acoustic events, we force the model to engage in perception processing rather than relying on shallow shortcuts or language priors.
3. **Perception Density Reward:** To stabilize the RL process, we introduce a novel perception density reward. This mechanism dynamically monitors the ratio of perception tokens in the generated CoT. It penalizes both under-perception and over-perception, thereby guiding the model toward a balanced reasoning trajectory.

Extensive experiments on MMAU, MMAU-Pro, and MMAR demonstrate that Audio-ReaPer sets a new state-of-the-art for audio reasoning. On the challenging MMAR benchmark, our 7B model achieves 68.5%, significantly outperforming the

previous best open-source model, and surpassing proprietary models like GPT-4o Audio and Gemini 2.5 Flash. Moreover, Audio-ReaPer matches the performance of sophisticated, high-latency agentic pipelines in an efficient end-to-end format. These results confirm that explicit perceptual grounding is the key to unlocking the internal reasoning potential of audio reasoning models.

2 Preliminaries: Common Practice for Reasoning Post-Training

Current state-of-the-art approaches for training language or multimodal reasoning models typically follow a two-stage paradigm: SFT for cold-starting the reasoning capabilities, followed by RLVR to further refine the policy.

SFT with CoT Data. To elicit the initial reasoning capabilities with longer outputs, the cold-start phase fine-tune the pre-trained model π_θ on human-annotated or synthetic question-answer pairs. Given an multimodal input x and a target reasoning chain $y = (y_1, \dots, y_L)$, The objective is a negative log-likelihood loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{SFT}}} \left[\sum_{t=1}^L \log \pi_\theta(y_t | x, y_{<t}) \right] \quad (1)$$

GRPO with Verifiable Rewards. For the RL stage, GRPO (Guo et al., 2025) is widely adopted due to its efficiency in eliminating the value network. For each query q , a group of outputs $\{o_i\}_{i=1}^G$ is sampled from the old policy $\pi_{\theta_{\text{old}}}$. In the context of reasoning tasks, the optimization is driven by a strictly verifiable reward function $R(o_i)$, which ensures both the exactness of the solution and compliance with the structural requirements:

$$R(o_i) = \alpha r_{\text{acc}}(o_i) + \beta r_{\text{fmt}}(o_i) \quad (2)$$

where r_{acc} is a binary reward indicating whether the final answer matches the ground truth, and r_{fmt} verifies adherence to the specific reasoning format, such as pre-defined “think” and “answer” tags. α and β are hyper-parameters to compute the reward. The optimization relies on normalizing the rewards within the group to compute the advantage. Specifically, the advantage is computed as:

$$A_i = \frac{R(o_i) - \text{mean}(R(o_j)_{j=1}^G)}{\text{std}(R(o_j)_{j=1}^G) + \epsilon} \quad (3)$$

The policy is then updated using the clipped surrogate objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left(\min \left(\rho_{i,t} A_i, \text{clip}(\rho_{i,t}, 1 - \epsilon, 1 + \epsilon) A_i \right) - \beta \mathbb{D}_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}) \right) \right] \quad (4)$$

where $\rho_{i,t} = \frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}$ is the importance sampling ratio for the t -th token in the i -th rollout, and $|o_i|$ denotes the length of the rollout. This standard formulation serves as the baseline for our experiments in Section 6.2 under the condition of comparable data. We also introduced the related Large Audio Language Models (LALM) and Large Audio Reasoning Models (LARM) in the Appendix A.

3 Bottleneck Analysis

3.1 Error Pattern Analysis

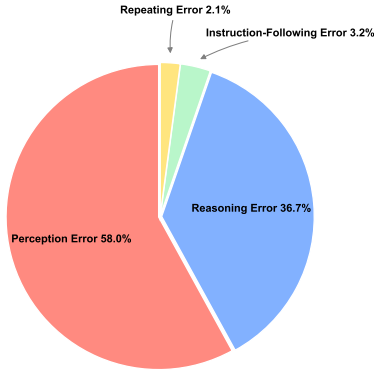


Figure 1: Qwen2.5-Omni reasoning error analysis of failure modes on the MMAR benchmark.

To understand the shortcomings of existing models, we adopt the error taxonomy proposed in MMAR (Ma et al., 2025b). Previous LARMs such as Audio-Reasoner (Xie et al., 2025), primarily based on Qwen2-Audio (Chu et al., 2024), indicated that perception errors accounted for the largest portion of failures. However, it remains an open question whether this bottleneck persists in stronger, state-of-the-art models. We apply the same taxonomic analysis to Qwen2.5-Omni (Xu et al., 2025a), a significantly more powerful model. As illustrated in Figure 1, despite the overall performance improvements, the error distribution exhibits a persistent pattern: perception errors remain the dominant failure mode. This suggests that training with a better language backbone does not automatically resolve the misalignment between audi-

tory features and semantic understanding required for complex reasoning tasks.

3.2 Oracle Injection Study

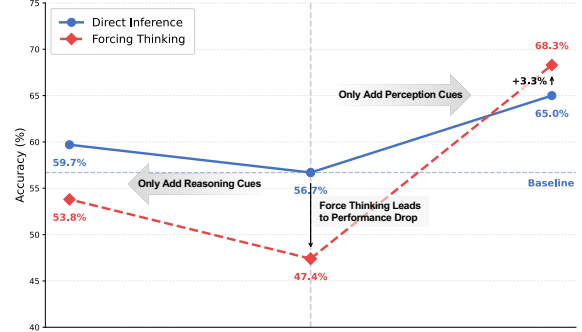


Figure 2: Oracle study with different cues to inject into Qwen2.5-Omni during reasoning.

The dominance of perception errors leads to a critical hypothesis:

Does the model lack the logical capacity to solve complex audio tasks, or is its reasoning engine simply hindered by flawed perception?

To answer this, we design an oracle injection study using the MMAR benchmark. We manually annotate the CoT for each sample and employ an LLM to decompose these thoughts into distinct **Perception Cues** (factual descriptions of acoustic events) and **Reasoning Cues** (logical steps required to derive the answer). We then inject these ground-truth cues into the prompt during inference to measure the performance upper bound when specific bottlenecks are removed. We evaluate Qwen2.5-Omni in two modes:

Direct Inference: The model directly predicts the answer without generating intermediate reasoning steps.

Forcing Thinking: The model is explicitly prompted to generate a reasoning path before outputting the final answer.

Figure 2 illustrates the results, revealing two pivotal insights regarding the interplay between perception and reasoning.

Insight 1: Perception cues are more critical than reasoning cues for improving audio reasoning.

In the Direct Inference mode, providing perception cues yields a significantly larger gain (+8.3%) compared to reasoning cues (+3.0%). This confirms that if the model is grounded with correct auditory information, it can often derive the correct answer using its internal knowledge, whereas correct reasoning logic is useless if applied to incorrect observations.

Insight 2: LALM has a intrinsic reasoning ability, but this potential need the cooperation of accurate perception to active.

Simply forcing the model to think degrades performance significantly (Instruct 56.70% → Thinking 47.40%). Analysis reveals that the reasoning mechanism acts as an “amplifier” for perceptual imperfections, turning sensory gaps into ungrounded hallucinations. Surprisingly, under the “Forcing Thinking” mode, the injection of perception cues not only recovers the performance drop but achieves the highest overall accuracy of 68.3%, outperforming the “Direct Inference + Perception Cues” setting (65.0%). This result is pivotal: it demonstrates that the model possesses strong reasoning capabilities unlocked by the CoT process, but this potential can only be realized when the perception bottleneck is resolved. In other words, accurate perception is the prerequisite that enables the intrinsic reasoning power of large audio language models.

This analysis demonstrates that the logical reasoning capability is largely intrinsic to the LLM backbone and is not the primary bottleneck. Once the perception bottleneck is resolved, the model’s internal reasoning power is effectively activated, transforming the “Thinking Paradox” into a significant performance advantage.

4 Audio-ReaPer

4.1 Training Paradigm

Agentic Data Generation. The scarcity of high-density perceptual data presents a primary obstacle for training LARMS. Existing audio captioning datasets typically feature brief, high-level descriptions (e.g., “a dog barking”), while standard Audio QA datasets often rely on simple pattern matching without requiring multi-hop perceptual synthesis. To overcome this, we utilize an agentic data generation pipeline (Ma et al., 2026) that synthesizes perception-dense data for both the cold-start and

reinforcement learning stages. As illustrated in Figure 3, the agent leverages an iterative “Query-Observation” mechanism to extract and refine fine-grained details. Instead of a single-pass description, the agent iteratively probes the observers to verify specific acoustic events (e.g., “Is the crashing sound caused by rain or a waterfall?”) and integrates the multi-turn evidence into a minimally hallucinatory, detailed caption. Appendix B shows the details of the data agent configuration.

Cross-Modal Consistency Filter. To ensure data reliability, we address the inherent ambiguity of audio perception, where, for example, the sound of heavy rain might be misidentified as a waterfall, or a dolphin’s call as a synthesized bird’s call. We employ a cross-modal verification strategy: we run the agentic data generation pipeline independently with both audio tools and video tools. By calculating the semantic consistency between the audio-derived and video-derived descriptions, we filter out instances with high perceptual discrepancy, retaining only the data where the auditory features are visually grounded and verifiable.

Perception-Enhanced SFT. In multimodal reasoning research, the cold-start phase typically follows the paradigm of DeepSeek-R1 (Guo et al., 2025), utilizing QA-based CoT data to standardize output style and encourage output length. However, we argue that this approach is suboptimal for the audio domain. Pure QA-based CoT data often suffers from sparse perception signals, and fine-tuning on such data can lead to perception degradation, where the model loses its ability to capture fine-grained acoustic details. Instead, we propose using **Perception-Enhanced Descriptive Data** for the cold start. Unlike standard QA samples, our data forces the model to maximize the likelihood of a dense perceptual description sequence before attending to any task. The SFT data contains rich, multi-perspective detailed captions (e.g., timbre, temporal relations, background events) generated by our agentic pipeline.

Perception-Focused RLVR. Building upon the strong perceptual foundation established by SFT, we construct data for RLVR. We employ an LLM to analyze the data agent trajectory and synthesize **Perception-Focused QA Data**. We enforce two criteria for these questions: (1) Multi-Cue Dependency: The answer must require synthesizing multiple correlated perceptual clues; and (2) Reasoning Depth: The question must involve logical deduction based on these observations. These strictly for-

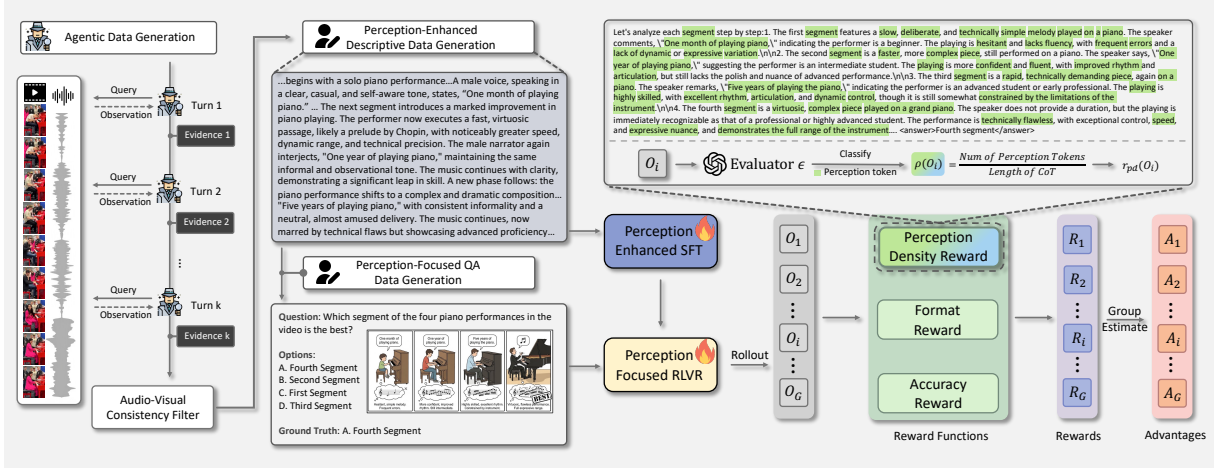


Figure 3: Illustration of the Audio-ReaPer. The framework includes Agentic Data Generation, Perception-Enhanced SFT, Perception-Focused RLVR, and Perception Density Reward to enhance perception of the model during reasoning post-training.

matted MCQs serve as the training ground for our subsequent RL stage. Building upon the capabilities established by the Perception-Enhanced SFT, we further conduct Perception-Focused RLVR to bridge the gap between perception and reasoning.

4.2 Perception Density Reward

During the reinforcement learning exploration phase, we observe that the model tends to collapse into two distinct degenerate modes when optimized solely on answer correctness:

- *Verbosity (Under-Perception)*: The model engages in “self-talk,” generating plausible-sounding reasoning paths that are entirely disconnected from the actual audio input, effectively ignoring perceptual evidence.
- *Stenography (Over-Perception)*: The model degenerates into length-exhausting descriptions, listing minute auditory cues without synthesizing them into a coherent logical chain.

To mitigate these issues, we propose the **Perception Density Reward**, a mechanism designed to constrain the reasoning trajectory within a balanced bound.

For a given audio query x , let $o_i = (w_1, \dots, w_T)$ be a generated reasoning chain consisting of T tokens. To quantify the information flow, we employ an evaluator $\mathcal{E}$ to dissect the sequence. The evaluator classifies each subsequence of o_i as either a *perception unit* (factual descriptions of acoustic events) or a *reasoning unit* (logical inferences and functional steps). Details of the evaluator design are provided in Appendix C. Let $\mathbb{I}_P(w_t)$ be an indicator function that equals 1 if token w_t belongs

to a perception unit and 0 otherwise. We define the **Perception Density Ratio** $\rho(o_i)$ as:

$$\rho(o_i) = \frac{1}{T} \sum_{t=1}^T \mathbb{I}_P(w_t) \quad (5)$$

We posit that a valid reasoning chain must maintain this density within a specific range $[\alpha, \beta]$ to ensure sufficient grounding without sacrificing reasoning capacity. The Perception Density Reward r_{pd} is defined as a gating function:

$$r_{pd}(o_i) = \begin{cases} 1, & \text{if } \alpha \leq \rho(o_i) \leq \beta \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Our preliminary experiments and token-level micro-dynamics analysis suggest that a strict gating mechanism provides a stronger regularization signal. Finally, the total reward $R(o_i)$ combines the verifiable task rewards r_{rlvr} (including correctness and format) with this perception density constraint:

$$R(o_i) = r_{rlvr}(o_i) + \lambda r_{pd}(o_i) \quad (7)$$

5 Experiment Setup

5.1 Training Dataset Construction

Our training data is derived from the AVQA dataset (Yang et al., 2022), a large-scale benchmark containing 57,015 unconstrained videos of daily activities. While AVQA provides a broad coverage of real-life scenarios, its original labels are insufficient for training deep reasoning capabilities due to sparse perceptual details. To address this, we implement a rigorous data refinement process using the agentic data generation pipeline

and a strict cross-modal verification protocol mentioned in Section 4.1. This filtering dramatically reduces the dataset from 57,015 to 17,923 samples, ensuring that every remaining instance possesses groundable, high-certainty acoustic signals. We then leverage this high-quality subset to generate task-specific training data: detailed captions for perception-enhanced SFT and multiple choice QA for perception-focused RLVR.

5.2 Training Setting

We employ Qwen2.5-Omni-7B as our backbone model. For both the Perception-Enhanced SFT and the Perception-Focused RLVR stages, we perform full-parameter fine-tuning. The specific hyperparameters are summarized in Appendix D.

5.3 Benchmarks

To ensure a rigorous evaluation audio reasoning capabilities, we select three representative benchmarks: MMAU (Sakshi et al., 2025), MMAU-Pro (Kumar et al., 2025), and MMAR (Ma et al., 2025b). Each benchmark targets a different level of cognitive processing, ranging from fundamental audio understanding to complex, multi-hop reasoning. Detailed dataset description are provided in Appendix E.

5.4 Baselines

To comprehensively evaluate the performance of Audio-ReaPer, we compare it against a wide range of state-of-the-art models, categorized into four groups: (1) **Proprietary Non-Thinking Models**; (2) **Proprietary Thinking Models**; (3) **Open-Source Non-Thinking Models**; and (4) **Open-Source Thinking Models**. Additionally, we include **agent-based systems**, as their multi-step execution naturally leverages external tools to provide strong perceptual grounding, offering a competitive reference for our end-to-end method. Specific model list is deferred to Appendix F. Furthermore, to simulate the current common practice in reasoning alignment, we construct a strong baseline. This baseline is trained using a standard pipeline: we utilize the high-quality information from our agentic data pipeline to prompt an LLM to generate Audio QA with CoT, followed by standard SFT and GRPO. This allows for a direct ablation study of our proposed perception-enhanced method against the standard “R1-style” reasoning training recipe in Section 6.2.

6 Experiment Results

6.1 Overall Performance Comparison

Table 1 presents a comprehensive evaluation of Audio-ReaPer against state-of-the-art proprietary and open-source models. The results demonstrate the effectiveness of our approach for training end-to-end audio reasoning models. We also show the data and training efficiency of our approach in Appendix G, and provide more comprehensive results in Appendix H.1.

Comparison with Backbone. Compared to our backbone model Qwen2.5-Omni, Audio-ReaPer achieves substantial improvements across all benchmarks. This confirms that our methods successfully prevent the “alignment tax” often observed in reasoning post-training, where specialized reasoning training degrades general capabilities.

Comparison with Open-Source Thinking Models. Audio-ReaPer significantly outperforms existing open-source reasoning models. When compared to R1-AQA, a pioneer implementation of the standard DeepSeek-R1 recipe without perception enhancement, the performance gap is evident (MMAU: 65.6% vs 76.4%). Furthermore, when compared to the previous SOTA model Audio-Flamingo 3 Thinking, our model leads by +8.4% on MMAR (60.1% → 68.5%). This empirical evidence strongly validates our core hypothesis that explicitly grounding the reasoning chain via perception density is essential for unlocking audio CoT reasoning.

Comparison with Agentic Methods. Remarkably, our end-to-end model achieves performance parity with sophisticated agent-based systems such as AudioGenie-Reasoner and AudioToolAgent. It suggests that a properly trained end-to-end model can internalize the “multimodal tool-use” process including detecting, analyzing, and reasoning as latent cognitive steps, eliminating the latency and complexity overhead of external tool calls.

Closing the Gap with Proprietary Models. Notably, on the challenging MMAU-Pro, our model (60.7%) outperforms the proprietary GPT-4o Audio (58.3%) and Gemini 2.0 Flash (57.3%). Similarly, on MMAR, it surpasses GPT-4o Audio (63.5%) and approaches the performance of the thinking-enabled Qwen3-Omni Flash Thinking (68.1%).

6.2 Training Stability

A common criticism of reinforcement learning in reasoning tasks is its training instability, particu-

Table 1: Performance comparison of LALMs (Non-Thinking) and LARMs (Thinking) on MMAU, MMAU-Pro, and MMAR, reported by category (Sound/Music/Speech) and overall average. The average scores “Avg” also include mixed modalities (Sound-Music, Sound-Speech, Music-Speech, Sound-Music-Speech) for MMAU-Pro and MMAR benchmarks. The best-performing models in each category are highlighted in **bold**, and the second-best ones are underlined. All numbers are taken from original papers or official benchmarks. Models are listed in chronological order of release.

Models	Size	MMAU (%)				MMAU-Pro (%)				MMAR (%)			
		Sound	Music	Speech	Avg	Sound	Music	Speech	Avg	Sound	Music	Speech	Avg
Random Guess	-	26.7	24.6	26.7	26.0	28.3	26.1	29.4	27.6	29.4	25.9	31.5	29.3
Proprietary Non-Thinking Models													
GPT-4o mini Audio	-	50.8	39.2	69.1	53.0	40.2	59.7	66.1	54.9	38.8	35.9	58.8	50.6
GPT-4o Audio	-	64.6	56.3	66.7	62.5	44.7	63.1	68.2	58.3	53.9	51.0	70.4	63.5
Gemini 2.0 Flash	-	71.2	65.3	75.1	70.5	48.4	56.9	69.5	57.3	61.2	51.0	72.1	65.6
Qwen3-Omni Flash Instruct	-	79.6	75.8	77.5	77.6	-	-	-	-	66.7	58.3	74.2	69.8
Proprietary Thinking Models													
Gemini 2.5 Flash	-	73.3	65.6	76.6	71.8	51.9	64.9	73.4	62.7	60.0	53.4	77.2	68.4
Audio-Thinker	-	77.5	70.4	73.4	73.7	-	-	-	-	67.3	53.9	64.3	65.3
Qwen3-Omni Flash Thinking	-	78.1	71.3	80.2	76.5	-	-	-	-	65.5	51.9	75.5	68.1
AudioGenie-Reasoner	Agent	75.1	66.2	76.6	72.6	-	-	-	-	49.7	43.3	69.2	58.9
AudioToolAgent	Agent	73.6	69.2	79.6	74.1	33.1	63.5	73.7	56.2	61.8	51.9	77.6	68.8
SAR-LM	Agent	73.3	65.0	82.3	73.5	-	-	-	-	52.7	56.3	81.0	69.3
Open-Source Non-Thinking Models													
Qwen2-Audio-Instruct	7B	55.0	51.0	42.0	49.2	-	-	-	-	33.3	24.3	32.3	30.0
SALMONN	7B	-	-	-	-	32.2	44.9	38.3	38.8	30.9	29.6	34.4	32.8
SALMONN	13B	41.0	34.8	25.5	33.7	43.6	47.2	37.3	43.4	30.3	31.1	34.7	33.2
Qwen2.5-Omni	7B	78.1	65.9	70.6	71.5	47.6	61.5	57.4	55.7	58.8	40.8	59.9	56.7
Audio-Flamingo 3	7B	<u>79.6</u>	<u>74.0</u>	66.4	73.3	55.9	<u>61.7</u>	<u>58.8</u>	<u>58.3</u>	-	-	-	58.5
Open-Source Thinking Models													
Audio-CoT	7B	61.9	56.3	55.3	57.8	-	-	-	-	35.8	25.2	34.0	31.3
Audio-Reasoner	7B	67.9	69.2	66.1	67.7	34.2	50.1	44.0	42.9	43.6	33.5	33.0	36.8
R1-AQA	7B	68.8	64.4	63.7	65.6	47.9	31.9	33.7	37.8	-	-	-	-
AudSemThinker	7B	69.7	69.2	61.3	66.7	-	-	-	-	-	-	-	-
Audio-Flamingo 3 Thinking	7B	79.9	76.6	66.4	<u>74.3</u>	-	-	-	-	-	-	-	<u>60.1</u>
AudioToolAgent-Open	Agent	78.1	69.7	75.1	74.2	42.8	64.4	67.9	58.0	59.4	45.6	67.3	61.7
Audio-ReaPer (ours)	7B	78.4	71.6	79.3	76.4	<u>52.6</u>	66.1	64.3	60.7	66.1	57.3	72.5	68.5

larly when reward signals are sparse. We compare Audio-ReaPer against the baseline mentioned in Section 5.4, which utilizes the same base model and data source but employs standard CoT training followed by standard GRPO. We analyze the validation accuracy curves across different training steps to verify the robustness of our approach. As visualized in Figure 4, Audio-ReaPer demonstrates a stabilized, monotonic ascent, and offers a more data-efficient and robust training paradigm than standard reasoning alignment.

6.3 Ablation Experiments

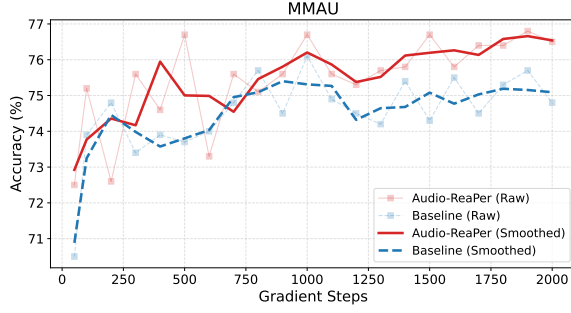
To verify the individual contributions of each component in Audio-ReaPer, we conduct ablation studies by removing key modules from the full pipeline, summarized in Table 2.

Impact of Perception-Enhanced SFT (Row a). Replacing our descriptive cold-start data with standard QA-based CoT data derived from the

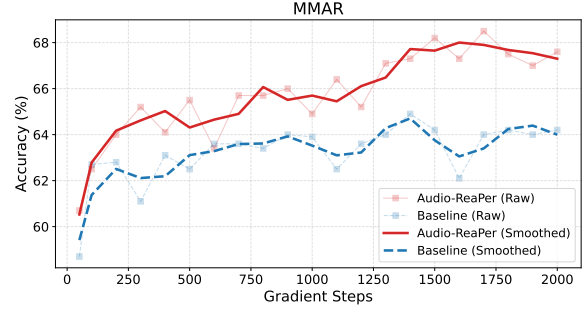
same source leads to a degradation, particularly on MMAR (68.5% $\rightarrow$ 65.2%). This suggests that without first establishing a dense perceptual mapping via detailed captioning, the model’s reasoning chain lacks a solid foundation, leading to ungrounded reasoning steps that are less effective for complex deduction.

Impact of Perception-Focused RLVR (Row b). Substituting our agentially generated multi-cue MCQs with the original, simpler AVQA questions results in the most significant performance drop (68.5% $\rightarrow$ 62.6% on MMAR). The original AVQA questions often rely on single-step pattern matching (e.g., "Is the dog barking?"). Training on such shallow tasks fails to cooperate with the exploration of RL, confirming that reasoning depth in the RL data is one of the drivers for unlocking the model’s potential.

Impact of Perception Density Reward (Row c). Removing the perception density constraint (op-



(a) MMAU accuracy curves across training steps.



(b) MMAR accuracy curves across training steps.

Figure 4: Training dynamics on the MMAU and MMAR benchmarks. Raw points are shown in light color and Savitzky–Golay smoothed curves in dark color.

timizing solely on $r_{\text{acc}} + r_{\text{fmt}}$) causes the MMAR score to decline to 67.0%. While the model still learns from the high-quality data, the absence of the density constraint makes it susceptible to the modes described in Section 6.2. The density reward effectively squeezes out the last mile of performance by regularizing the reasoning trajectory.

Table 2: Ablation study for different methods we propose to enhance perception on the MMAU and MMAR benchmarks.

Method Variant	MMAU (%)	MMAR (%)
Audio-ReaPer	76.4	68.5
(a) w/o Perception-Enhanced SFT	74.5	65.2
(b) w/o Perception-Focused RLVR	74.1	62.6
(c) w/o Perception Density Reward	75.0	67.0

We also verify our pipeline’s robustness by substituting the closed-source evaluator with open-source models. Comprehensive experimental results and analysis are provided in Appendix H.2.

6.4 The Goldilocks Zone for Perception Density Interval

To determine the optimal boundaries $[\alpha, \beta]$ for the Perception Density Reward, we conduct a grid search over the density interval parameters. Figure 5 visualizes the MMAR accuracy as a function of the *Interval Midpoint* ($\frac{\alpha+\beta}{2}$) and *Interval Length* ($\beta - \alpha$). As shown in the heatmap, the performance forms a clear **bell curve distribution**. The global optimum is achieved at a midpoint of 0.4 and a length of 0.4, corresponding to the interval $[\alpha = 0.2, \beta = 0.6]$. This setting yields the peak MMAR accuracy of 68.5%. This confirms that the Perception Density Reward is effective to enforce a balanced information flow, avoiding the extremes of ungrounded speculation and vacuous description.

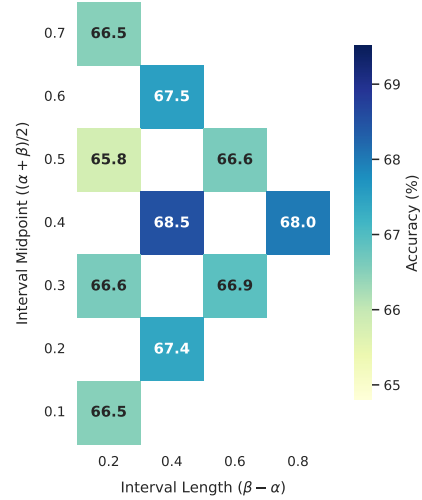


Figure 5: Impact of Perception Density Interval settings on MMAR accuracy. The heatmap axes represent the Interval Midpoint (vertical) and Interval Length (horizontal).

7 Conclusion

In this work, we identified that the primary bottleneck impeding the reasoning capabilities of LALMs is not logical deduction, but perceptual precision. Through rigorous error analysis and oracle studies, we demonstrated that "thinking" without accurate "hearing" leads to hallucinated reasoning chains. To address this, we introduced Audio-ReaPer, a comprehensive post-training framework that explicitly fortifies perception in the design of SFT data, RL data, and RLVR reward. Our approach achieves state-of-the-art performance across diverse benchmarks, including MMAU, MMAU-Pro, and MMAR. The success of our method highlights a critical insight for the multimodal reasoning community: the optimization of abstract reasoning objectives often induces perceptual degradation. By enforcing a perception constraint during post-training, we effectively prevent this alignment tax.

Limitations

While our post-training framework, Audio-ReaPer, successfully prevents the alignment tax by enforcing a perception constraint, its current implementation is confined to the post-training stage, treating high-fidelity perception as a post-hoc alignment target rather than an innate capability. Looking forward, we posit that the principle of “Perception Density” should extend beyond post-training. Integrating large-scale detailed multimodal captions into the pre-training or mid-training stages represents a promising direction. This would allow future models to internalize high-fidelity perception as an innate capability, paving the way for truly general-purpose, native audio reasoning models.

Potential Risks

This paper presents work whose goal is to advance the field of multimodal learning, specifically focusing on improving the reliability and reasoning depth of Large Audio Language Models (LALMs). On the positive side, our proposed method, Audio-ReaPer, explicitly mitigates the “hallucination” problem in audio reasoning by grounding generation with perception. This contributes to the development of safer and more trustworthy AI systems, which are crucial for applications such as assistive technologies for the hearing impaired, automated safety monitoring, and complex acoustic environment analysis. However, we acknowledge that advanced capabilities in fine-grained audio perception and reasoning could potentially be repurposed for enhanced surveillance or unauthorized audio analysis. We emphasize the importance of deploying such technologies with robust privacy safeguards. Furthermore, our research relies on publicly available datasets, and we do not foresee specific data privacy issues within the scope of this work.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62671373 and No. U23B2018), the Science and Technology Innovation (STI) 2030-Major Project (2022ZD0208700), and Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2024CSJGG1100).

8 Generative AI Use Disclosure

AI-assisted tools were used exclusively for linguistic editing and proofreading of this manuscript. The

original ideas and content were developed entirely by the authors.

References

- Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2023. Beats: Audio pre-training with acoustic tokenizers. *Proc. 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*.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.
- 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*.
- Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *ICML*.
- 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. *Proc. NeurIPS*.
- Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023. Joint audio and speech understanding. In *Proc. ASRU*.
- Google. 2025. Gemini 2.0 flash. <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash>.
- 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. *Proc. Nature*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

- Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *Proc. ICML*.
- Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeon-gon 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*.
- 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*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *Proc. COLM*.
- 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. *Proc. ASRU*.
- 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. *Proc. NeurIPS*.
- Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, and 1 others. 2026. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *Proc. ICLR*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.
- Yan Rong, Chenxing Li, Dong Yu, and Li Liu. 2025. Audiogenie-reasoner: A training-free multi-agent framework for coarse-to-fine audio deep reasoning. *Proc. ICASSP*.
- 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? *Proc. ASRU*.
- 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. *Proc. ICLR*.
- Termeh Taheri, Yinghao Ma, and Emmanouil Benetos. 2025. Sar-lm: Symbolic audio reasoning with large language models. *arXiv preprint arXiv:2511.06483*.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. *Proc. ICLR*.
- Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, and 1 others. 2025. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*.
- Gijs Wijnngaard, Elia Formisano, and Michel Dumontier. 2025a. Audiotoolagent: An agentic framework for audio-language models. *arXiv preprint arXiv:2510.02995*.
- Gijs Wijnngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. 2025b. Audsemthinker: Enhancing audio-language models through reasoning over semantics of sound. *Proc. NeurIPS*.
- Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. 2025. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*.
- Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. *Proc. EMNLP*.
- 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, Xinfu Zhu, and 1 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.
- Dongchao Yang, Songxiang Liu, Disong Wang, Yuanyuan Wang, Guanglu Wan, and Helen Meng. 2025. Omni-autothink: Adaptive multimodal reasoning via reinforcement learning. *arXiv preprint arXiv:2512.03783*.
- 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 *Proc. ACM Multimedia*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Proc. NeurIPS*.

A Related Moels

A.1 Large Audio Language Models (LALM)

Recent years have witnessed a paradigm shift in audio processing, moving from specialized task-specific models to general-purpose LALMs. The standard architecture typically connects a pre-trained audio encoder (e.g., Whisper (Radford et al., 2023), BEATs (Chen et al., 2023)) with an LLM via a light-weight projector. Early pioneers like LTU (Gong et al., 2023) and SALMONN (Tang et al., 2024) demonstrated that aligning audio features with text instructions enables models to perform general audio captioning and question answering. Subsequently, Qwen-Audio series (Chu et al., 2023, 2024) and Audio-Flamingo series (Kong et al., 2024; Ghosh et al., 2025) scaled up this paradigm, enabling multi-task handling across speech, sound, and music through massive multi-task fine-tuning. Despite their success in short-form understanding tasks such as speech recognition and audio event detection, these models fundamentally lack the capability to perform complex, multi-step logical deduction. They often struggle with queries requiring the synthesis of multiple acoustic clues, limiting their utility in rigorous reasoning scenarios.

A.2 Large Audio Reasoning Models (LARM)

To address the reasoning limitations of LALMs, the community has begun exploring LARMs that incorporate explicit CoT methodologies. Audio-CoT (Ma et al., 2025a) pioneered this direction by introducing CoT to LALMs via in-context prompting, though it found that gains were often limited to difficult tasks. Subsequently, Audio-Reasoner (Xie et al., 2025) and Audio-Flamingo 3 (Goel et al., 2025) scaled this approach by fine-tuning LALMs on large-scale synthetic CoT datasets, such as CoTA (Xie et al., 2025) and AF-Think (Goel et al., 2025), enabling the models to generate structured reasoning paths. More recently, inspired by DeepSeek-R1 (Guo et al., 2025), there is a growing interest in applying RLVR to enhance reasoning. R1-AQA (Li et al., 2025) and Omni-R1 (Rouditchenko et al., 2025) attempted to directly apply GRPO to audio question answering; however, they observed that the explicit “thinking” mode often failed to outperform non-thinking baselines. To overcome this, AudSemThinker (Wijngaard et al., 2025b), Qwen3-Omni Thinking (Xu et al., 2025b), and Step-Audio-R1 (Tian et al., 2025) utilize high-quality cot-start data followed by modified GRPO algorithms. Other models, such as Audio-Thinker (Wu et al., 2025) and Omni-AutoThink (Yang et al., 2025) apply adaptive thinking to improve the intelligence of reasoning during RL post-training. These approaches successfully enable models to generate valid, modality-grounded reasoning chains, achieving state-of-the-art performance across complex audio benchmarks.

B Agentic Data Pipeline Details

The agentic data generation pipeline acts as a sophisticated multi-model ensemble designed to synthesize, verify, and refine fine-grained multimodal captions. By integrating the complementary strengths of proprietary and open-weight models, the pipeline ensures both semantic depth and temporal precision. We employ GPT-4o as the central controller. Its advanced text ability is responsible for formulating investigative queries, synthesizing evidence, and performing final hallucinations checks. To process raw multimodal inputs, we utilize Gemini 2.5 Flash and GPT-4o Audio. These models serve as the sensory front-end, providing direct audio understanding and visual context interpretation. To enhance data production throughput, we leverage Qwen3-Omni Flash and Qwen3-Omni-Captioner. These models specialize in rapid caption generation and multimodal grounding, ensuring the scalability of our pipeline. Precise transcription and temporal localization are handled by GPT-4o Transcribe for speech recognition and Qwen3-ASR Flash, which is specifically utilized to provide accurate timestamp annotations. The generation process is modeled as a multi-turn dialogue between the “Detective” and the “Observer”. We set the maximum number of interaction steps to 10. To ensure efficiency, the agent is configured with an early-stopping mechanism, allowing the investigative process to conclude immediately once the information converges and a comprehensive description is constructed.

C Evaluator Design for Perception Density

To compute the Perception Density Reward r_{pd} , we require a robust mechanism to classify whether each segment of the generated CoT is a perception unit. Given the complexity of natural language, rule-based keyword matching is insufficient. Therefore, we implement an LLM-based Evaluator utilizing GPT-4o to perform semantic classification.

Evaluator System Prompt

You are an expert in multimodal data analysis, specialized in deconstructing complex Chains of Thought (CoT) into raw perceptual information and logical reasoning processes.

Evaluator Prompt

Task

Analyze the provided Chain of Thought (CoT) text and extract **all** words and phrases related to **perception**.

- Output the result in a JSON format with a single key: `perception_words`.
- The list must follow the **sequential order** of the original text.
- **Duplicates are required:** If a perceptual word or fact appears multiple times in the text, it must be listed every time it appears.

Definition of Perception

`perception_words` include:

- Direct sensory inputs (audio, visual, tactile, etc.).
- Physical characteristics or literal observations (e.g., "high-pitched", "stone", "well", "300 meters").
- Literal quotes of what was heard or seen (e.g., specific spoken sentences).
- Physical actions described in the scene (e.g., "thrown", "decays", "bouncing").

Constraints

- **Exhaustive Extraction:** Do not omit any sensory or physical detail.
- **Sequence-wise:** Maintain the exact order as they appear in the input.
- **No Summarization:** Do not merge multiple mentions into one; keep them as individual entries in the list.
- **Exclude Reasoning:** Do not include logical connectors (e.g., "therefore"), inferences (e.g., "this suggests"), or structural markers (e.g., "Step 1").

Example

Input

Let's break down the reasoning step by step:

1. The audio clip begins with a male voice speaking in Mandarin Chinese, instructing the listener to "Please calculate the depth of this well based on the echo." This sets the context for a scientific or educational demonstration.
2. Immediately after the instruction, a stone is thrown into the well. The sound of the stone is a sharp, high-pitched "tink," which quickly decays into a long, complex echo. The echo is not a single, short sound but a layered, reverberant decay that persists for several seconds.
3. The echo's decay is long and complex, with multiple overlapping reflections.

Output

```
{{ "perception_words": [ "audio clip", "male voice", "Mandarin Chinese", "Please calculate the depth of this well based on the echo", "stone", "thrown", "into the well", "sound of the stone", "sharp", "high-pitched", "tink", "long", "complex echo", "echo", "single, short sound", "layered", "reverberant decay", "persists for several seconds", "echo's decay", "long", "complex", "multiple overlapping reflections" ] }}
```

New Input {cot}

D Training Details

Table 3 lists the training hyperparameters for Perception-Enhanced SFT, and Table 4 lists the training hyperparameters for Perception-Focused RLVR. For the RL stage, to accommodate the sparsity of our binary perception density reward, we adopt a simplified advantage estimation strategy that removes the variance normalization term, following recent findings (Liu et al., 2025). To ensure fine-grained control over the generation trajectory and handle varying sequence lengths, we also adopt the token-level policy gradient objective (Yu et al., 2025).

Table 3: Hyperparameters for Perception-Enhanced SFT.

Parameter	Value
GPU Usage	$8 \times \text{A100 80GB}$
Batch Size per GPU	2
Gradient Accumulation	4
Peak Learning Rate	$5\text{e-}6$
LR Scheduler	Linear
Optimizer	AdamW
Training Epochs	2
Training Time	8 hours

Table 4: Hyperparameters for Perception-Focused RLVR.

Parameter	Value
<i>General Optimization</i>	
GPU Usage	$8 \times \text{A100 80GB}$
Batch Size per GPU	1
Gradient Accumulation	8
Peak Learning Rate	$1\text{e-}6$
LR Scheduler	Cosine
Optimizer	AdamW
Training Steps	2000
Training Time	22 hours
<i>GRPO Specifics</i>	
Group Size (G)	8
Updates per Rollout (μ_{epochs})	1
Clip Coefficient (ϵ)	$[0.2, 0.28]$
KL Coefficient (β_{KL})	0
<i>Generation & Rollout</i>	
Sampling Temperature	1
Top- p	0.9
<i>Perception Density Reward Settings</i>	
Density Lower Bound (α)	0.2
Density Upper Bound (β)	0.6
Regularization Weight (λ)	1

E Benchmark Details

MMAU MMAU (Massive Multi-Task Audio Understanding) (Sakshi et al., 2025) serves as our testbed for evaluating the model’s fundamental perception and basic reasoning capabilities. It comprises approximately 10k audio clips paired with human-annotated question-answer pairs. The benchmark covers 27 distinct audio understanding skills distributed across sound, music and speech domains. As the Test set is not open-sourced, we used the Test-mini set with 1k MCQ for our experiments. The Test-mini set remains a reliable evaluation set as it keeps the same task distribution as the full Test set.

MMAU-Pro MMAU-Pro (Kumar et al., 2025) represents a significant scaling up in difficulty and scope, designed to evaluate “Audio General Intelligence”. It features longer audio contexts (up to 10 minutes), open-ended generation tasks, and a wider array of skills including spatial awareness and conversational agents. Since MMAU-Pro contains specific tasks (e.g., spatial audio, conversational voice) that diverge from the scope of audio reasoning, we evaluate on a rigorously selected subset aligned with the MMAR taxonomy, including sound, music and speech for single modality, and sound-music, sound-speech, music-speech, and sound-music-speech for mixed-modalities, resulting in 3,548 test samples in total.

MMAR MMAR (Massive Multi-disciplinary Audio Reasoning) (Ma et al., 2025b) is a benchmark

specifically designed to assess the “System 2” reasoning capabilities of LALMs. Unlike MMAU, which focuses on information retrieval, MMAR requires the model to perform multi-step logical deductions based on acoustic evidence. It consists of 1,000 expert-annotated audio and QA samples. MMAR evaluates models across four hierarchical levels of abstraction: signal level, perception level, semantic level, and cultural level. Crucially, MMAR introduces mixed-modality tasks (e.g., sound-music, sound-music-speech), requiring the model to disentangle and reason over overlapping audio streams.

F Baseline Details

This section provides a detailed model list of the comprehensive baselines introduced in Section 5.4:

- **Proprietary Non-Thinking Models:** GPT-4o mini Audio (Hurst et al., 2024), GPT-4o Audio (Hurst et al., 2024), Gemini 2.0 Flash (Google, 2025), and Qwen3-Omni Flash Instruct (Xu et al., 2025b).
- **Proprietary Thinking Models:** Gemini 2.5 Flash (Comanici et al., 2025), Audio-Thinker (Wu et al., 2025), and Qwen3-Omni Flash Thinking (Xu et al., 2025b).
- **Open-Source Non-Thinking Models:** Qwen2-Audio-Instruct (Chu et al., 2024), SALMONN (Tang et al., 2024), Qwen2.5-Omni (Xu et al., 2025a), and Audio-Flamingo 3 (Goel et al., 2025).
- **Open-Source Thinking Models:** Audio-CoT (Ma et al., 2025a), Audio-Reasoner (Xie et al., 2025), R1-AQA (Li et al., 2025), AudSemThinker (Wijngaard et al., 2025b), and Audio-Flamingo 3 Thinking (Goel et al., 2025).
- **Agent-Based Systems:** AudioGenie-Reasoner (Rong et al., 2025), AudioToolAgent (Wijngaard et al., 2025a), and SAR-LM (Taheri et al., 2025) (which rely on proprietary LLMs/LALMs), as well as AudioToolAgent-Open (Wijngaard et al., 2025a) (which uses open-source tools). These are included because their sequential tool-use pipelines explicitly enhance multi-step perceptual grounding.

G Connecting Perception and Reasoning via Efficient Training.

Table 5 highlights the data and training efficiency of our approach. Other approaches rely on massive amounts of QA-based SFT data to implicitly learn audio reasoning. However, this often results in suboptimal performance because the model attempts to jump directly to logical deduction without a solid perceptual foundation. In contrast, Audio-ReaPer explicitly bridges this gap through Perception-Enhanced SFT and Perception-Focused RLVR. This perception-first strategy allows Audio-ReaPer to achieve state-of-the-art performance with roughly 10% of the data volume required by previous baselines, demonstrating that explicitly connecting perception to reasoning is far more effective than brute-force scaling of QA pairs.

Table 5: Comparison of base models, post-training paradigms, data efficiency, and performance on the MMAR audio reasoning benchmark among Large Audio Reasoning Models. “Org.” represents “Original” while “Mod.” represents “Modified”.

Model	Base Model	Post-Training Paradigm	SFT Data	RL Data	MMAR (%)
Audio-CoT	Qwen2-Audio-Instruct	Prompt Engineering	-	-	31.3
Audio-Reasoner	Qwen2-Audio-Instruct	SFT	1.2M	-	36.8
R1-AQA	Qwen2.5-Omni	Org. GRPO	-	38K	-
AudSemThinker	Qwen2.5-Omni	SFT + Org. GRPO	797K	140K	-
Audio-Flamingo 3 Thinking	Audio-Flamingo 3 Base	SFT	250K	-	60.1
Audio-ReaPer (Ours)	Qwen2.5-Omni	SFT + Mod. GRPO	18K	18K	68.5

H More Results

H.1 Full Results

Table 6 and Table 7 report the full results including sound, music and speech for single modality, and sound-music, sound-speech, music-speech, and sound-music-speech for mixed-modalities.

Table 6: Performance comparison of LALMs (Non-Thinking) and LARMs (Thinking) on MMAU-Pro, reported by category (Sound/Music/Speech/Sound-Music/Sound-Speech/Music-Speech/Sound-Music-Speech) and overall average. All numbers are taken from original papers or official benchmarks. Models are listed in chronological order of release.

Models	Size	Single Modality (%)			Mixed Modalities (%)				Avg (%)
		Sound	Music	Speech	Sound-Music	Sound-Speech	Music-Speech	Sound-Music-Speech	
Random Guess	-	28.3	26.1	29.4	24.2	30.5	25.2	14.8	27.6
Proprietary Non-Thinking Models									
GPT-4o mini Audio	-	40.2	59.7	66.1	35.3	55.9	42.2	42.8	54.9
GPT-4o Audio	-	44.7	63.1	68.2	40.4	62.5	43.5	57.1	58.3
Gemini 2.0 Flash	-	48.4	56.9	69.5	39.6	55.9	57.6	42.8	57.3
Proprietary Thinking Models									
Gemini 2.5 Flash	-	51.9	64.9	73.4	42.8	61.3	58.7	42.8	62.7
AudioToolAgent	Agent	33.1	63.5	73.7	26.0	54.6	50.0	57.1	56.2
Open-Source Non-Thinking Models									
SALMONN	7B	32.2	44.9	38.3	22.0	28.4	34.8	28.6	38.8
SALMONN	13B	43.6	47.2	37.3	28.0	38.4	47.8	42.8	43.4
Qwen2.5-Omni	7B	47.6	61.5	57.4	40.0	60.2	53.2	28.5	55.7
Audio-Flamingo 3	7B	55.9	61.7	58.8	40.0	47.7	41.3	57.1	58.3
Open-Source Thinking Models									
Audio-Reasoner	7B	34.2	50.1	44.0	26.0	43.2	36.9	28.6	42.9
R1-AQA	7B	47.9	31.9	33.7	32.0	20.4	36.9	28.5	37.8
AudioToolAgent-Open	Agent	42.8	64.4	67.9	36.0	54.6	45.7	57.1	58.0
Audio-ReaPer (ours)	7B	52.6	66.1	64.3	40.0	56.8	43.5	57.1	60.7

H.2 Evaluator Robustness

To address potential concerns regarding the pipeline’s reliance on closed-source models for the RLVR Evaluator, we conducted additional scaling experiments by substituting the default GPT-4o judge with open-source alternatives, specifically Qwen3-235B-A22B and Qwen3-8B, while keeping all other training and environment configurations strictly constant.

As illustrated in Table 8, the massive open-source Qwen3-235B-A22B achieves highly comparable performance to GPT-4o, even slightly outperforming it on MMAU. This demonstrates that our reinforcement learning pipeline does not suffer from proprietary lock-in and can be effectively deployed within a fully open-source ecosystem. On the other hand, the smaller Qwen3-8B model exhibits a noticeable performance degradation across all evaluation sets. We attribute this bottleneck to the extensive length and logical complexity of our evaluator prompt, which mandates rigorous structured output compliance and strict multi-criteria instruction following—capabilities that smaller parameter-scale models inherently struggle to stabilize during continuous evaluation loops. These scaling behaviors underscore that while high-capacity models are required for precise evaluation, our framework successfully scales to open-weights architectures.

I Statement For The Use of Large Language Models

AI-assisted tools were used exclusively for linguistic editing and proofreading of this manuscript. The original ideas and content were developed entirely by the authors.

Table 7: Performance comparison of LALMs (Non-Thinking) and LARMs (Thinking) on MMAR, reported by category (Sound/Music/Speech/Sound-Music/Sound-Speech/Music-Speech/Sound-Music-Speech) and overall average. All numbers are taken from original papers or official benchmarks. Models are listed in chronological order of release.

Models	Size	Single Modality (%)			Mixed Modalities (%)				Avg (%)
		Sound	Music	Speech	Sound-Music	Sound-Speech	Music-Speech	Sound-Music-Speech	
Random Guess	-	29.39	25.88	31.48	25.00	29.30	31.10	28.13	29.32
Proprietary Non-Thinking Models									
GPT-4o mini Audio	-	38.8	35.9	58.8	45.5	60.1	57.3	50.0	50.6
GPT-4o Audio	-	53.9	51.0	70.4	63.6	72.5	62.2	75.0	63.5
Gemini 2.0 Flash	-	61.2	51.0	72.1	82.8	72.5	65.9	70.8	65.6
Qwen3-Omni Flash Instruct	-	66.7	58.3	74.2	81.8	78.4	64.6	70.8	69.8
Proprietary Thinking Models									
Gemini 2.5 Flash	-	60.0	53.4	77.2	63.6	76.2	69.5	75.0	68.4
Audio-Thinker	-	67.3	53.9	64.3	72.7	71.6	73.2	66.7	65.3
Qwen3-Omni Flash Thinking	-	65.5	51.9	75.5	63.6	76.2	65.9	70.8	68.1
AudioGenie-Reasoner	Agent	49.7	43.3	69.2	45.5	64.5	65.3	59.1	58.9
AudioToolAgent	Agent	61.8	51.9	77.6	72.7	76.6	72.0	70.8	68.8
SAR-LM	Agent	52.7	56.3	81.0	-	-	-	-	69.3
Open-Source Non-Thinking Models									
Qwen2-Audio-Instruct	7B	33.3	24.3	32.3	9.1	31.2	30.5	25.0	30.0
SALMONN	7B	30.9	29.6	34.4	9.1	37.6	28.1	37.5	32.8
SALMONN	13B	30.3	31.1	34.7	9.1	34.9	35.4	41.7	33.2
Qwen2.5-Omni	7B	58.8	40.8	59.9	54.5	61.9	67.1	67.1	56.7
Audio-Flamingo 3	7B	-	-	-	-	-	-	-	58.5
Open-Source Thinking Models									
Audio-CoT	7B	35.8	25.2	34.0	9.1	30.7	30.5	37.5	31.3
Audio-Reasoner	7B	43.6	33.5	33.0	45.5	42.7	31.7	25.0	36.8
R1-AQA	7B	55.8	37.4	49.0	9.1	50.0	50.0	50.0	47.6
Audio-Flamingo 3 Thinking	7B	-	-	-	-	-	-	-	60.1
AudioToolAgent-Open	Agent	59.4	45.6	67.3	54.6	70.6	59.8	70.8	61.7
Audio-ReaPer (ours)	7B	66.1	57.3	72.5	81.8	75.2	68.3	66.7	68.5

Table 8: Performance comparison on downstream benchmarks when training with proprietary versus open-source models as the RLVR evaluator.

Evaluator Model	MMAU (%)	MMAU-Pro (%)	MMAR (%)
GPT-4o	76.4	60.7	68.5
Qwen3-235B-A22B	76.7	60.7	68.1
Qwen3-8B	75.0	56.2	66.6