When Scaling Fails: Mitigating Audio Perception Decay of LALMs via Multi-Step Perception-Aware Reasoning

Anonymous ACL submission

Abstract

Test-Time Scaling has shown notable efficacy in addressing complex problems through scaling inference compute. However, within Large Audio-Language Models (LALMs), an unintuitive phenomenon exists: post-training models for structured reasoning trajectories results in marginal or even negative gains compared to training for direct answering. To investigate it, we introduce **CAFE**, an evaluation framework designed to quantify audio reasoning errors. Evaluation results reveal that LALMs encounter a critical bottleneck: **Audio Perception Decay**, which compromises reasoning performance. To address it, we propose **MPAR**^{2}, a strategy that encourages a step-by-step reasoning process with rich perceptual awareness. With reinforcement learning, MPAR^2 improves performance on CAFE from 31.74% to 63.51% and effectively mitigates audio perception decay, concurrently enhancing reasoning ability to achieve a strong 74.59% accuracy on the MMAU benchmark. Further analysis shows that MPAR^2 optimizes audio attention and enhances reasoning token budget efficiency.

1 Introduction

Test-Time Scaling (Snell et al., 2024) has emerged as a dominant paradigm that enables models to solve complex problems by scaling inference compute, typically realized through extended chains of thought (CoT) (Guo et al., 2025; Wei et al., 2022; Brown et al., 2024). Inspired by this success, recent works have extended similar strategies to Large Audio-Language Models (LALMs) through reinforcement learning (RL), yielding notable gains on complex audio tasks that demand deep audio understanding and reasoning (Tian et al., 2025; Wu et al., 2025). However, a marked exception persists: optimizing models for structured reasoning trajectories via RL results in marginal or even negative gains compared to direct-answering tuning. As illustrated in Figure 1 on the MMAU benchmark (Sakshi et al., 2024), direct-answer RL baselines, which are trained via implicit prompts, outperform CoT-based variants like R1-AQA (Li et al., 2025) and SARI (Wen et al., 2025), where explicit reasoning prompts are utilized to generate <thinking> content. Meanwhile, Audio-Thinker (Wu et al., 2025) achieves comparable results. This phenomenon differs from the expectation of scaling that increasing inference compute yields predictable performance improvement. In light of this, we raise a question: *what causes the ineffectiveness of extended reasoning in complex audio tasks?*

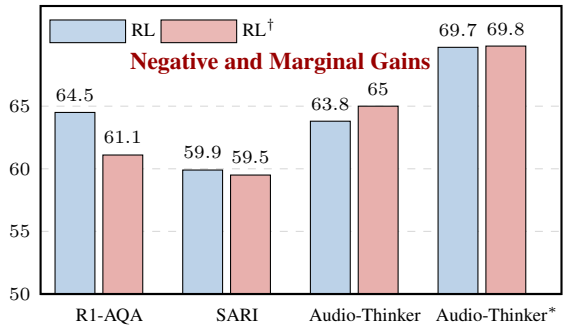


Figure 1: Comparison of direct-answer RL training and RL with explicit reasoning prompts (denoted by RL†). Three RL variants (R1-AQA, Audio-Thinker, and SARI) are tested on the MMAU benchmark. Here, * denotes Qwen2.5-Omni as the base model; otherwise, Qwen2-Audio is used. More details of the prompts used and response cases are presented in Appendix B.5.

Recent research has attempted to tackle similar challenges. Step-Audio-R1 (Tian et al., 2025) attributes this issue to an over-reliance on textual reasoning, exploring modality-grounded frameworks to mitigate the gap. Meanwhile, an analysis in MMAR (Ma et al., 2025b) reveals audio event perception and utilization errors are the dominant factor. However, despite these efforts, a systematic empirical investigation of this problem is still rare.

To fill this void, we introduce **C**omprehensive **A**udio **F**idelity **E**valuation framework (**CAFE**), which is designed to quantify audio event fidelity

during reasoning process. Leveraging the "LLM-as-a-Judge" paradigm (Zheng et al., 2023), CAFE extracts and counts audio events from reasoning processes to quantify perception and utilization errors. Experimental results reveal that current audio models lack the ability to stably anchor acoustic cues across long generated sequences, *audio perception decay* prevents extended reasoning from delivering its intended positive impact. As the reasoning chain extends, the model gradually omits or hallucinates events, with the accumulation of such errors degrading the accuracy of logical deductions.

To address this limitation, we propose **MPAR**^{2}, a **M**ulti-step **P**erception-**A**ware **R**easoning and **R**eview strategy designed to mitigate perception decay, thereby enhancing reasoning capability. Extending the power of CoT, MPAR^2 establishes a "think-while-listening" paradigm. It follows a structured pipeline: initially enforcing a fine-grained scan for explicit audio perception, followed by stepwise perception-rich question decomposition, and concluding with a review process to check overall reasoning details and rationality. To instantiate this strategy, we adopt a two-stage training scheme. First, supervised cold-start training endues formatted reasoning ability. In the second stage, RL further optimizes the model for audio perception and reasoning. Experiments show MPAR^2 achieves 74.59% on MMAU (origin) test-mini and 60.32% on MMAR. Under the CAFE evaluation, MPAR^2 achieves leading perception and utilization accuracy. Further analysis shows that MPAR^2 reinforces audio attention during reasoning and dynamically adjusts reasoning budget based on question complexity. In summary, our contributions are:

- We introduce **CAFE**, an evaluation framework that provides a powerful quantitative tool for assessing perception and reasoning quality during the reasoning process of audio models.
- We present the first systematic analysis of *audio perception decay* in LALMs, where longer reasoning chains lead to degraded audio perception, resulting in marginal or even negative gains from complex CoT reasoning.
- We propose **MPAR**^{2}, an RL-based strategy designed to enhance both audio event perception and utilization. Experiments show that MPAR^2 achieves strong perception and reasoning accuracy, while reinforcing audio attention and enhancing reasoning token budget efficiency.

2 Related Work

Large Audio-Language Models As large language models (LLMs) continue to advance, models that integrate multiple modalities have gradually gained growing focus. These multimodal LLMs are capable of jointly modeling and reasoning over various inputs, including audio. Recent LALMs, such as Qwen2-Audio (Chu et al., 2024), Qwen2.5-Omni (Xu et al., 2025), DeSTA2.5-Audio (Lu et al., 2025), and MiMo-Audio (Xiaomi, 2025), have demonstrated notable effectiveness in understanding and processing acoustic information.

Large Audio Reasoning Models Recent research has increasingly focused on enhancing LALMs with reasoning capability. While early efforts like Audio-CoT (Ma et al., 2025a) and Audio-Reasoner (Xie et al., 2025) introduced CoT via prompting or supervised fine-tuning (SFT), but achieved limited gains and failed to unlock the reasoning potential of audio models. Subsequently, reinforcement learning has recently pushed boundaries. Methods such as R1-AQA (Li et al., 2025) and SARI (Wen et al., 2025) utilize RL to optimize structured CoT reasoning, whereas Omni-R1 (Rouditchenko et al., 2025) adopts a direct-answer strategy that surprisingly surpasses CoT. However, results on the variants of R1-AQA and SARI reveal an unintuitive behavior in audio test-time scaling. Although other strategies like Audio-Thinker (Wu et al., 2025) explore an adaptive thinking mode, Step-Audio-R1 (Tian et al., 2025) applies modality-grounded frameworks to further improve reasoning capability, and CESAR (Fan et al., 2025) first analyses it and applies fine-grained reward method.

Adaptive Reasoning and Difficulty Awareness The success of reasoning models such as OpenAI o1 (Jaech et al., 2024) demonstrates the effectiveness of test-time scaling, while also highlighting the practical need to reduce computational redundancy. Prior work shows that LLMs implicitly encode problem difficulty in their hidden representations (Lee et al., 2025), enabling adaptive reasoning termination. Building on this insight, S-GRPO (Dai et al., 2025) employs reinforcement learning with decaying rewards to encourage early correct exits, while data-centric methods (Waheed et al., 2025) further improve token efficiency by distilling difficulty-aware reasoning length through compressed CoT trajectories.

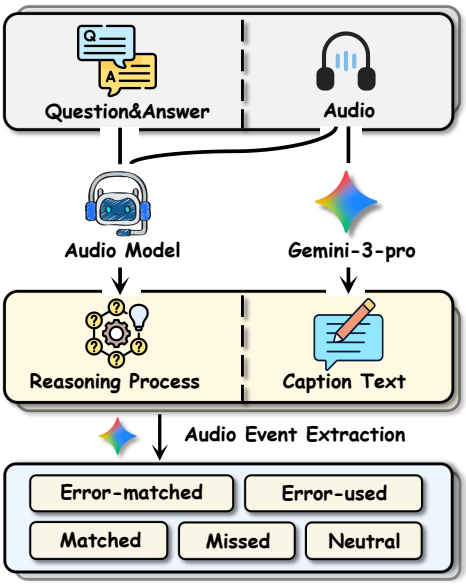


Figure 2: Overview of the CAFE framework evaluation pipeline. First, audio model generates a reasoning process based on the audio-QA input. Meanwhile, the audio is captioned by Gemini-3-Pro. Finally, Gemini-3-Pro utilizes the inputs, the reasoning process, and the captions to extract different audio events.

3 Reasoning-Time Audio Event Perception and Utilization Probing

In this section, we define the task of reasoning-time audio event perception and utilization probing and introduce the CAFE framework. Subsequently, we present and analyze evaluation results across LALMs, Large Audio Reasoning Models (LARMs), and commercial models.

Task Formulation Given ground-truth audio content and Question-Answer (QA) pairs, the reasoning-time audio event perception and utilization probing task aims to quantify the identification and utilization of audio events during model reasoning, employing metrics such as perception accuracy, utilization accuracy, and omission rate. We leverage a powerful textual LLM to process the probing inputs. This approach enables robust generalization across semantic variations, while ensuring the precise extraction of audio events strictly required for the question-response context. Furthermore, it enables the quantification of audio fidelity throughout statistically analyzing the extracted audio events.

3.1 Overview

To address this task, we introduce the CAFE framework. Formally, employing the probing textual model $\mathcal{P}$, the audio event extraction process can be formulated as:

$$\mathcal{E} = \mathcal{P}(O, Q, A, C) \tag{1}$$

where O represents the model's reasoning output generated by the model, Q and A denote the input question and the target answer, and C is the ground truth audio caption. The same abbreviations are used throughout the subsequent formulas. The resulting output $\mathcal{E}$ consists of categorized audio events with labels E_i, where aspect $E_i \in \{Matched, Hallucinated, Misused, Missed, Neutral\}$, defined as follows: (1) *Matched*: existing audio events correctly identified and utilized during the reasoning process to address the question. (2) *Hallucinated*: events that are hallucinatory or misidentified during reasoning; (3) *Misused*: existing events that are logically misused, leading to incorrect conclusions; (4) *Missed*: existing events required for the answer but omitted during reasoning; and (5) *Neutral*: existing events mentioned but irrelevant to the reasoning outcome. Furthermore, we extract and count these categorized events to derive specific metrics below.

Metric Formulation We define four distinct metrics based on the extracted event categories. Let $N_{(\cdot)}$ denote the count of events for each category. We first define the prediction space N_{pred} and the target space N_{tgt} as follows:

$$\begin{aligned} N_{\text{pred}} &= N_{\text{mat}} + N_{\text{hal}} + N_{\text{misuse}} + N_{\text{neu}}, \\ N_{\text{tgt}} &= N_{\text{mat}} + N_{\text{miss}} \end{aligned} \tag{2}$$

where N_{pred} denotes the total number of audio events in model reasoning process, and N_{tgt} represents the complete set of ground-truth events required to answer the question. Based on these, we derive the following metrics:

- *Perception Accuracy* (Acc_{per}) is defined as the ratio of correctly identified events related to the QA task, formulated as $Acc_{\text{per}} = \frac{N_{\text{mat}}}{N_{\text{pred}}}$. Acc_{per} is aimed at perception in reasoning rather than simple audio caption.
- *Perception Error Rate* (Err_{per}) quantifies the rate of hallucinatory or misidentified events via $Err_{\text{per}} = \frac{N_{\text{hal}}}{N_{\text{pred}}}$.
- *Usage Error Rate* (Err_{use}) assesses events that are correctly perceived but logically misused, which can be formulated as $Err_{\text{use}} = \frac{N_{\text{misuse}}}{N_{\text{pred}}}$.
- *Omission Error Rate* (Err_{omit}) quantifies the events required for the answer that the model failed to capture, defined as $Err_{\text{omit}} = \frac{N_{\text{miss}}}{N_{\text{tgt}}}$.

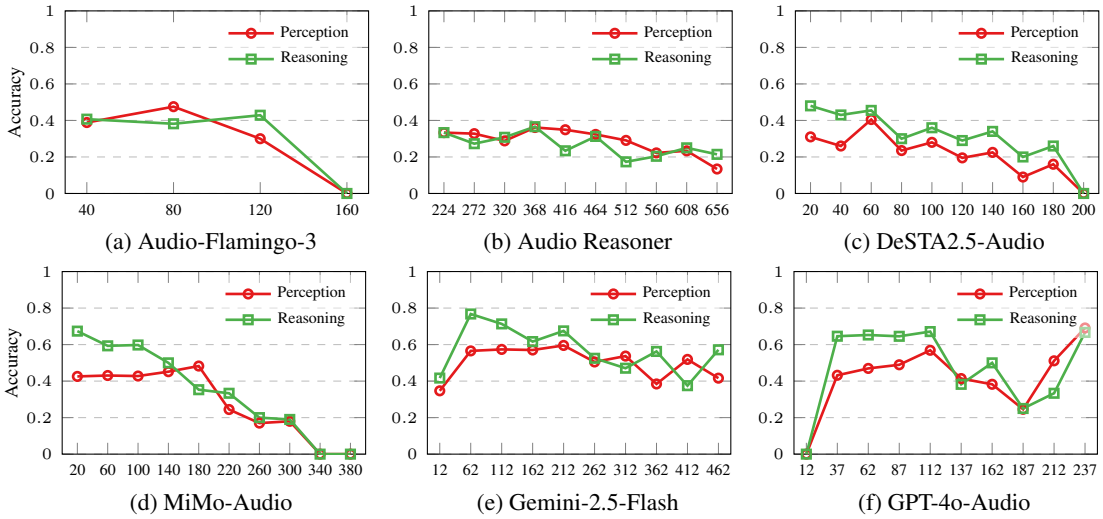


Figure 3: **Reasoning Accuracy** (accuracy on the evaluated MMAR benchmarks) and CAFE's perception accuracy across reasoning token lengths. The x-axis represents the mean of token length intervals, and the y-axis indicates the accuracy. Reasoning accuracy aligns closely with perception accuracy. Specifically in the (a)–(d), both metrics decline as reasoning tokens increase, which we term *Audio Perception Decay*.

In particular, Acc_{per} offers a holistic assessment of the audio model's perceptual fidelity during reasoning. This metric serves as the quantitative basis for Figure 3, where we track perceptual performance as reasoning length increases, ultimately revealing the phenomenon of audio perception decay.

Framework details Within CAFE, Gemini-3-Pro generates captions for the benchmark audio and serves as the probing model $\mathcal{P}$. Replacing it with GPT-5 and using semantically equivalent prompts in an ablation study leads to performance variations within ±2.1%. The overall workflow of the framework is shown in Figure 2. Additional prompts and details are provided in Appendix A.

3.2 Evaluation

Experimental Setup For models that lack explicit reasoning output, we employ designed prompts to elicit reasoning. A unified prompt is used for the remaining models. All prompts, details and verification are presented in Appendix B.

3.3 Evaluation Results

To analyze the relationship between reasoning length and both perception and reasoning accuracy, we selected two representative models from each category: LALMs, LARMs, and commercial models, as shown in Figure 3. Table 1 presents the overall results, including the MPAR^2 method, which is detailed in Section 4. Based on these results, we conduct analysis and identify two findings.

Audio Perception Decay Across Extended Reasoning As illustrated in Figure 3 (a)-(d), open-source LARMs and LALMs exhibit a clear downward trend: both perception and reasoning accuracy decline as reasoning length increases, eventually dropping to zero. Pearson (Benesty et al., 2009) analysis confirms a robust positive correlation between these two metrics across (a)-(d) groups, yielding coefficients ranging from 0.65 to 0.91, where the majority reach statistical significance with $p < 0.05$. In contrast, commercial models (e)-(f) maintain stable and superior performance without exhibiting decay trend. Thus, the observed variations in these metrics show that perception struggles to maintain a robust connection with the expanding reasoning chain. Instead, perception accuracy degrades as reasoning length increases, which we term **"Audio Perception Decay"**, leading to an accumulation of errors that drives the degradation of reasoning capability.

Weak Perception during Reasoning As shown in Table 1, most audio models perform poorly on CAFE for MMAR. Notably, open-source models rarely surpass the threshold of 40% Acc_{per}, and even advanced systems like Gemini-2.5-Flash reach only 54.81%, with the majority falling below 30%. However, curated RL strategies prove to significantly enhance capabilities. Specifically, Omni-R1 (51.21%) substantially outperforms its Qwen2.5-Omni baseline (31.74%), while Step-Audio-R1.1 (56.18%) and Step-Audio-R1 (53.08%) similarly achieve remarkable perfor-

Model	Acc_{per} ↑	Err_{per} ↓	Err_{use} ↓	Err_{omit} ↓	$MMAR_{acc}$ ↑
Large Audio Language Models					
MiMo-Audio	42.79	27.70	22.40	48.84	59.87
Qwen2.5-Omni-7B	31.74	29.61	28.69	57.37	55.20
DeSTA2.5-Audio†	23.19	33.95	35.57	69.32	41.60
Phi-4-Multimodal	20.43	36.75	30.32	73.41	39.80
Qwen2-Audio-Instruct	8.27	56.46	28.58	81.01	29.90
Omni-R1	51.21	28.79	12.90	42.65	62.10
Large Audio Reasoning Models					
Step-Audio-R1.1	56.18	23.77	11.25	21.49	**67.50**
Step-Audio-R1	53.08	24.82	12.42	**6.77**	67.40
Audio-Flamingo-3†	41.39	26.03	23.18	52.87	56.40
Audio Reasoner	27.44	25.55	32.58	55.72	36.80
Commercial Models					
Gemini-2.5-Flash	54.81	19.27	15.96	33.24	66.30
GPT-4o-Audio	48.68	**17.20**	21.71	38.05	63.80
MPAR^2-7B	**63.51**	23.14	**7.74**	30.59	60.32

Table 1: Results of various audio models on the MMAR benchmark within the CAFE framework. † denotes $MMAR_{acc}$ evaluated by GPT-5 judge. All results are averaged over three inference runs. The best performance in each group is in **bold** and the second best one is underlined. The remaining table settings are the same.

mance. For instance, the Step-Audio-R1 series exhibits no clear bottleneck across the four metrics, whereas other models show marked disparities, underscoring that reasoning capability is critically sensitive to even partial perceptual deficits. These findings reveal that *effective audio reasoning requires a strong grounding in audio perception*, rather than merely producing well-formatted but semantically disconnected outputs.

4 Methodology

To address the perception decay of reasoning over extended sequences, we propose a novel reasoning paradigm, **M**ulti-step **P**erception **A**ware **R**easoning and **R**eview strategy (**MPAR^2**), which is designed to structurally reinforce audio perception and reasoning while maintaining compatibility with inference-time scaling strategies. In this section, we present the MPAR^2 CoT structure, rewards design and describe its two-stage training procedure. The overview of MPAR^2 is in Figure 4.

4.1 CoT structure of MPAR^2 Framework

We begin by defining the reasoning process under the MPAR^2 paradigm. Each reasoning chain follows three stages: (1) *Perception*: Explicitly describe the audio events related to the QA task. Detail their inter-relationships, strictly following the time sequence; (2) *Reasoning*: Dynamically decompose the original question into a structured sequence of sub-questions, adjusting the sub-question steps based on the question's difficulty. Then, solve the sub-questions using audio-perception details and reasoning, until the core inquiry is fully resolved. (3) *Review*: Double-check the fidelity of the audio events used during reasoning, and ensure the logical consistency. Make necessary corrections to the final answer if errors are detected.

4.2 Multi Rewards Designs

To both refine perception quality and promote adaptive reasoning budgets within the MPAR^2 paradigm, we leverage Group Relative Policy Optimization (GRPO) (Shao et al., 2024). This optimization process is steered by three key generative rewards and format reward, which jointly work to enhance both reasoning and perceptual fidelity. Specifically, we utilize a powerful textual LLM S as a score model to obtain process-level rewards, guided by designed prompts. The prompt designs and more details are provided in Appendix F.

Perception Reward Given the severe perceptual limitations and audio perception decay during reasoning, MPAR^2 begins the reasoning process with an explicit audio perception step. It incorporates a perception reward to ensure high-precision estimation, as shown in Figure 4. Formally, using the scoring model S, the reward is defined as:

$$R_{\text{perception}} = S(O_{\text{perception}}, Q, A, C) \qquad (3)$$

where $O_{\text{perception}}$ represents the perception part of CoT . Specifically, given the ground truth caption with QA, we prompt Qwen3-32B as the model S to assess the generated perception. The model outputs a score $R_{\text{perception}} \in [0, 1]$ as reward with a specific focus on the audio events essential for solving the problem, evaluating them based on faithfulness, completeness, and consistency.

Stepwise Perception-Reasoning Reward Prior studies (Marjanović et al., 2025; Lee and Hockenmaier, 2025) demonstrate that structured reasoning trajectories, such as problem decomposition, significantly enhance complex problem-solving. Meanwhile, existing approaches such as Audio-Thinker (Wu et al., 2025) primarily rewards the overall reasoning process in a coarse manner, but lacks fine-grained guidance over reasoning quality and explicit emphasis on audio perception. Inspired by these findings, we introduce Stepwise Perception-Reasoning Reward (SPR Reward) to reinforce structured reasoning.

SPR Reward uses two key criteria: *stepwise validity* and *chain consistency*. Specifically, stepwise validity evaluates the event completeness and logical soundness of each sub-QA. It aims for high

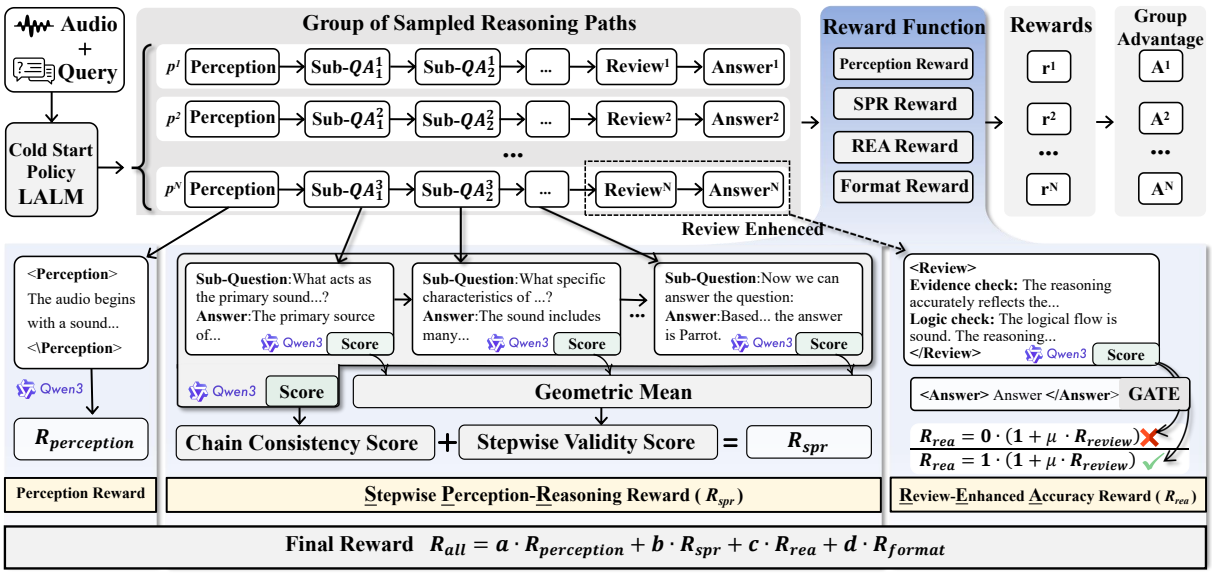


Figure 4: An illustration of MPAR^2 training pipeline. The top of the figure outlines the overall GRPO training setup. The lower section provides a view of the reward design. Cold start policy model is derived from SFT-trained model.

efficiency and direct resolution within the current step, ensuring that every question posed is both critical and perception-rich. We calculate these step scores using the geometric mean (McAlister, 1879), which is described in Eq. 5, a method chosen to enforce uniform quality across the entire process by strictly penalizing any single step failure.

Moving beyond individual steps, chain consistency evaluates the logic of the entire reasoning path. Crucially, this mechanism encourages the model to adaptively adjust its sub-question decomposition based on question difficulty, thereby controlling the reasoning budget. It requires making effective use of detailed audio descriptions and analysis to establish a clear reasoning, while enforcing a strict order in which audio analysis precedes and supports final problem-solving.

With these criteria, we define SPR Reward as:

$$R_{\text{spr}} = \theta \cdot \bar{S}_{\text{sub-reason}} + (1-\theta) \cdot S_{\text{all-reason}} \quad (4)$$

$\bar{S}_{\text{sub-reason}}$ and $S_{\text{all-reason}}$ are calculated by:

$$\bar{S}_{\text{sub-reason}} = \left(\prod_{i=1}^{n} S\big(O^{(i)}_{\text{sub-reason}}, Q, A, C\big) \right)^{\frac{1}{n}} \quad (5)$$

$$S_{\text{all-reason}} = S\big(O_{\text{all-reason}}, Q, A, C\big) \quad (6)$$

where $\theta \in [0, 1]$ allocates the reward weights between individual reasoning steps and entire reasoning path, which is set to 0.7. $O^{(i)}_{\text{sub-reason}}$ denotes the i-th sub-reasoning step, and $O_{\text{all-reason}}$ is the overall reasoning content. Notably, for stepwise validity and chain consistency evaluation, in addition to the reasoning content, we also input the audio caption C, the question Q, and the ground-truth answer A, thereby improving the evaluation quality.

Review-Enhanced Accuracy Reward Current explicit reasoning in audio models is inherently one-pass. Once the model generates reasoning steps, it continues without looking back to re-examine the alignment between the audio content and its text. Consequently, if an initial perceptual error occurs (e.g., misidentifying a "cat" as a "dog"), subsequent reasoning steps merely reinforce this error rather than correcting it. This lack of an "auditory feedback loop" necessitates a mechanism for backward verification. To address this, we introduce the Review-Enhanced Accuracy Reward (REA Reward), which integrates this double-check with a direct reward for the correctness of the final answer. We calculate it:

$$R_{\text{rea}} = R_{\text{acc}} \cdot (1 + \mu \cdot R_{\text{review}}) \quad (7)$$

where we set μ to 0.5, and R_{review} is defined as:

$$R_{\text{review}} = S(O_{\text{review}}, O_{\text{all-reason}}, Q, A, C) \quad (8)$$

where O_{review} denotes the review part of model output. R_{acc} is defined as a binary reward, taking the value of 1 if the generated answer aligns with the ground truth, and 0 otherwise. Notably, we treat R_{acc} as a gating factor to prevent that review process is rewarded but the answer is still wrong, as shown in Eq. 7. This ensures that the review reward is only granted when the answer is correct.

Model	Size	MMAU (Test-mini)				MMAR			
		Sound↑	Music↑	Speech↑	Avg↑	Sound↑	Music↑	Speech↑	Avg↑
Large Audio Language Models									
MiMo-Audio	7B	**80.78**	70.96	63.06	71.60	58.79	43.69	60.54	59.87
Qwen2.5-Omni	7B	69.40	66.80	61.60	65.90	53.94	42.23	58.84	55.20
DeSTA2.5-Audio†	8B	58.26	44.61	58.26	53.70	53.74	24.76	29.70	41.60
Phi-4-Multimodal	5.6B	57.96	51.80	48.35	52.70	32.12	30.58	45.58	39.80
Qwen2-Audio-Instruct	8.4B	49.85	52.40	42.94	48.40	33.33	24.27	32.31	29.90
Omni-R1	8.4B	73.60	**74.30**	66.10	71.30	67.30	51.50	64.30	62.10
Kimi-Audio	11.3B	61.68	73.27	60.66	65.00	48.48	29.13	42.52	40.60
Large Audio Reasoning Models									
Audio-Flamingo-3†	7B	79.58	72.46	60.36	70.80	58.26	44.61	58.26	56.40
Audio Reasoner*	8.4B	60.06	64.30	60.70	61.71	43.64	33.50	32.99	36.80
SARI*	7B	72.75	67.22	61.26	67.08	-	-	-	-
R1-AQA*	8.4B	68.77	64.37	63.66	65.60	-	-	-	-
Audio-Thinker*	7B	77.48	70.36	73.37	73.70	**68.32**	**53.88**	64.29	65.30
Step-Audio-R1	33B	72.97	61.68	74.47	69.70	57.58	45.63	77.89	67.40
Step-Audio-R1.1	33B	72.67	72.46	**77.18**	74.10	60.00	45.15	**79.25**	**67.50**
Commercial Models									
Gemini-2.5-Flash	-	67.56	61.48	63.76	64.27	53.33	50.97	75.85	66.30
GPT-4o-Audio	-	60.36	57.19	67.27	61.60	54.55	48.54	69.39	63.80
MPAR^2-7B	7B	77.98	71.30	74.47	**74.59**	62.40	44.30	64.26	60.32
w/o $R_{\text{perception}}$	7B	75.84	72.90	71.34	73.36	60.10	42.60	63.82	58.84
w/o R_{spr}	7B	77.74	70.68	73.78	74.08	61.90	43.80	63.16	59.62
w/o R_{spr} and $R_{\text{perception}}$	7B	77.19	70.66	71.43	73.11	59.60	42.40	62.60	58.20
w/o R_{spr}, $R_{\text{perception}}$ and R_{rea} ‡	7B	77.54	67.32	73.27	72.81	58.90	41.80	61.86	57.52

Table 2: Accuracy (%) on MMAU (original) and MMAR benchmark. † denotes results evaluated by GPT-5. * means results are reproduced, while all other results are averaged over three inference runs. ‡ means the ablation setting with a simple accuracy reward.

Format Reward We define a format reward R_{format} of 1 for outputs complying with the MPAR^2 structure illustrated in Figure 4, and 0 otherwise.

Overall Reward After assigning weights to each component, the final reward R_{all} is defined as:

$$R_{\text{all}} = \alpha \cdot R_{\text{perception}} + \beta \cdot R_{\text{spr}} + \gamma \cdot R_{\text{rea}} + \delta \cdot R_{\text{format}} \quad (9)$$

where we set α, β, γ, δ to 1.5, 1.0, 1.5, and 0.1.

5 Experiments

5.1 Model Setup

In our experiments, we adopt the Qwen2.5-Omni (Xu et al., 2025) as the base model. After SFT, we perform GRPO training on 4 H100 GPUs. The reward model is served via vLLM. More detailed training settings are provided in the Appendix E.1.

Data Construction We construct our dataset from AVQA (Yang et al., 2022) by generating captions with Gemini-2.5-Pro and synthesizing four-dimensional reasoning QA pairs via Qwen3-32B[1] (Yang et al., 2025). We filter for consistency and retain only samples with intermediate difficulty, resulting in 46,544 candidates $Data_{RL}$. A balanced subset of 5,000 is selected for SFT preparation. Qwen3-32B generates CoT paths for this subset, yielding 4,600 high-quality instances $Data_{SFT}$ for Stage-1 supervised fine-tuning (SFT), while the remaining 41,544 samples are reserved for Stage-2 RL. Detailed generation procedures, prompts, and four dimensions' case studies are presented in Appendix E.

[1] https://huggingface.co/Qwen/Qwen3-32B

5.2 Main Results

We primarily evaluate model performance on the MMAU and MMAR benchmarks, and further conduct reasoning analysis under the CAFE framework, with benchmark details provided in Appendix C. The MMAU-v05.15.25 and 3B model's results in Appendix D. From the experimental results in Tables 1, 2 and Figure 5, we conclude:

MPAR^2 Facilitates Effective Perception Refinement As demonstrated in Table 1 and Figure 5, MPAR^2 effectively addresses the challenge of perception decay and boosts perceptual capabilities. Specifically, in Figure 5, MPAR^2 successfully mitigates the trend of perception decay, maintaining both perception and reasoning accuracy at consistently high and stable levels. In Table 1, MPAR^2 achieves leading performance among the evaluated models in terms of perception accuracy Acc_{per} and event utilization error Err_{use}, highlighting the potential of our method. Meanwhile, to gain deeper insight into perceptual behavior, we conduct more detailed analysis in Appendix D.2 and D.3.

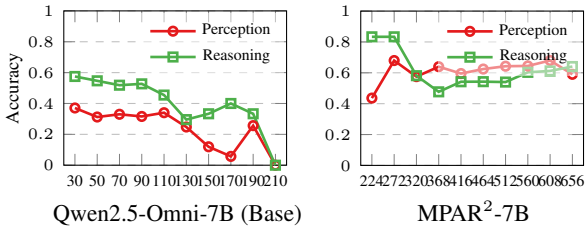


Figure 5: Comparison of accuracy across token lengths.

MPAR^2 Stably Improves Perception and Reasoning Capabilities As shown in Table 2, compared to the Qwen2.5-Omni-7B baseline, MPAR^2 achieves substantial performance gains. Specifically, MPAR^2 boosts accuracy on MMAU from 65.90% to 74.59% and on MMAR from 55.20% to 60.32%. Notably, stable improvements are observed across distinct audio domains, which supports our core insight: *enhanced perception during reasoning can suppress audio perception decay and boost reasoning performance*. However, the gain on MMAR is smaller than on MMAU, and MPAR^2 remains behind Step-Audio-R1, while achieving higher perception accuracy on CAFE. We attribute these differences to two factors that jointly affect the reasoning performance of LALMs: *audio perception capacity during reasoning* and *inherent logical reasoning capacity*. MPAR^2's main goal is to strengthen perceptual anchoring throughout the entire reasoning process, with greater emphasis on the former factor. For stronger logical ability, we will further explore data and ways to enhance text-based reasoning ability in future work.

6 Analysis

Ablation Study We show the impact of different reward ablations in Table 2, demonstrating that each mechanism contributes positively to MPAR^2. First, relying solely on format and simple accuracy rewards already yields substantial gains of 6.91% on MMAU and 2.32% on MMAR over the baseline, validating the effectiveness of the MPAR^2 pattern. Second, incorporating the review-enhanced accuracy mechanism leads to further improvements across all metrics, underscoring the robustness and efficacy of the rethinking process. Finally, the integration of perception and multi-step reasoning rewards specifically boosts perception capabilities, elevating MMAU performance from 73.36% and 74.08% to 74.59%. Notably, the perception reward contributes a relatively larger gain, highlighting the critical role of perception in audio reasoning. Meanwhile, More ablations on model size and method are provided in Appendix D.4 and D.5.

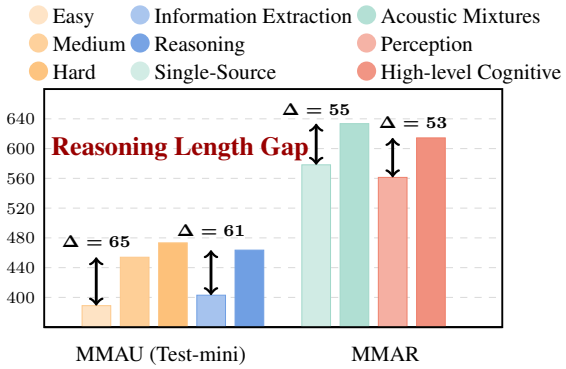


Figure 6: Token length of MPAR^2-7B by task category on MMAU and MMAR. The y-axis reports tokens.

Adaptive Reasoning Budget and Latency Trade-off As illustrated in Figure 6, we summarized the reasoning token lengths of MPAR^2 across different aspects on benchmarks. Notably, since MMAR focuses on deep reasoning problems, the model exhibits a longer overall output length compared to MMAU. Beyond this dataset-level observation, we find that straightforward inputs, such as easy-level questions or basic information extraction tasks, lead to shorter reasoning, whereas more complex problems result in longer responses. This behavior suggests that our MPAR^2 successfully enables the model to dynamically adjust its reasoning budget rather than following a fixed long CoT pattern. Specifically, it reduces the reasoning budget for simple tasks while extending it for complex scenarios, the model reduces the unnecessary overhead introduced by CoT in simple audio QA tasks, achieving a better alignment between reasoning budget and task complexity.

7 Conclusion

In this work, we first investigate why RL for structured reasoning yields marginal or even negative gains over RL for direct answering. Based on the results obtained from the proposed CAFE framework, we identify the critical bottleneck: as reasoning scales, reasoning capability suffers from audio perception decay. To address it, we propose MPAR^2, which successfully mitigates perception decay, improves perceptual performance during reasoning, further enhancing overall reasoning ability. Additionally, in-depth analysis shows that MPAR^2 optimizes audio attention during reasoning and dynamically adjusts reasoning budget allocation.

Limitations

Model Size Limit Due to GPU constraints, our experiments are limited to models with a maximum size of 7B. Future work will extend training to larger-scale models to further validate the method's effectiveness.

Depth of Logical Reasoning Our current dataset and method primarily focus on perceptual enhancement during reasoning. To further bolster the reasoning capabilities of LALMs, future work will explore more challenging QA data and advanced methodologies aimed at deepening inherently logical reasoning proficiency.

Reasoning Depth vs. Real-Time Latency In spoken question answering, test-time scaling style reasoning improves performance but usually increases inference time. This creates a fundamental trade-off between reasoning depth and real-time latency, since dialogue systems require fast, natural, and responsive interaction. Traditional sequential pipelines follow a listen, think, then speak paradigm, which preserves reasoning quality but introduces substantial delay. Although MPAR^2 improves computational efficiency, it still does not fully meet real-time latency demands. Future work will focus on more efficient reasoning and dynamic scheduling of reasoning and response generation to better balance accuracy and responsiveness.

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*.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *arXiv preprint arXiv:2404.01077*.

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*.

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*.

Muzhi Dai, Chenxu Yang, and Qingyi Si. 2025. S-grpo: Early exit via reinforcement learning in reasoning models. *arXiv preprint arXiv:2505.07686*.

Jiajun Fan, Roger Ren, Jingyuan Li, Rahul Pandey, Prashanth Gurunath Shivakumar, Ivan Bulyko, Ankur Gandhe, Ge Liu, and Yile Gu. 2025. Incentivizing consistent, effective and scalable reasoning capability in audio llms via reasoning process rewards. *arXiv preprint arXiv:2510.20867*.

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. *arXiv preprint arXiv:2507.08128*.

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*.

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*.

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*.

Jinu Lee and Julia Hockenmaier. 2025. Evaluating step-by-step reasoning traces: A survey. *arXiv preprint arXiv:2502.12289*.

Sunbowen Lee, Qingyu Yin, Chak Tou Leong, Jialiang Zhang, Yicheng Gong, Shiwen Ni, Min Yang, and Xiaoyu Shen. 2025. Probing the difficulty perception mechanism of large language models. *arXiv preprint arXiv:2510.05969*.

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*.

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. *arXiv preprint arXiv:2505.13032*.

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, and 1 others. 2025. Deepseek-r1 thoughtology: Let's think about llm reasoning. *arXiv preprint arXiv:2504.07128*.

Donald McAlister. 1879. Xiii. the law of the geometric mean. *Proceedings of the Royal Society of London*, 29(196-199):367–376.

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. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

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

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

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*.

Abdul Waheed, Chancharik Mitra, Laurie Z Wang, Deva Ramanan, and Bhiksha Raj. 2025. Less is more tokens: Efficient math reasoning via difficulty-aware chain-of-thought distillation. *arXiv preprint arXiv:2509.05226*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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*.

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*.

LLM-Core-Team Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.

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

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

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 *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A CAFE Framework Prompts

In terms of prompt design, we have referred to efficient prompting methods (Chang et al., 2024). The following are the Gemini-3-Pro caption and audio event extraction prompts used by the CAFE framework.

A.1 Gemini-3-Pro Caption prompt

Please generate a detailed chronological description of the following audio clip.

Listen carefully to the sequence of events and describe the audio flow from beginning to end. Your caption should:
- Explicitly state the order of sounds using transitional phrases (e.g., The audio begins with..., Followed by..., Simultaneously..., As the sound fades...).
- Capture the subtle details of each sound event, including its duration and intensity changes.
- Distinguish between foreground events (main actions) and background noise (ambience).

Output the result as a cohesive narrative text without line breaks or bullet points.

A.2 Audio Event Extraction prompt

Role
You are an expert Audio Logic Consistency Evaluator. Your task is to evaluate a Model Reasoning Path against the Question, Correct Answer, and Ground Truth Caption.

Step 1: Analyze Requirements
1. Determine ‘required_events‘: The essential sounds from the [Ground Truth Audio Caption] needed to answer the [Question] correctly.
2. Identify ‘all_caption_events‘: All sounds actually present in the audio.

Step 2: Categorize Model Events
For every audio event mentioned in the [Model Reasoning Path], categorize it into ONE of the following 4 lists based on its Validity (Is it real?) and Usage (How did the model use it?):
1. matched_events (Correct & Necessary):
- The event is in ‘required_events‘.
- The model used it effectively to derive the answer.
2. error_matched (Fabrication / Misidentification):
- The event is NOT in ‘all_caption_events‘ (Hallucination).
- OR The model heard sound A (real) but identified it as sound B (fake/wrong), and B is not in the caption.
- These are False Positives regarding perception.
3. error_use (Distraction / Wrong Reasoning):
- The event exists in ‘all_caption_events‘ but is NOT in ‘required_events‘ (Irrelevant).
- CRITICAL CONDITION: The model activley used this irrelevant sound to support a conclusion, or the model was confused by it.
- Example: I hear a bird (irrelevant), so the answer must be ’Forest’. (When the answer is actually ’Park’ due to other sounds).
4. neutral_events (Harmless Mention / Valid Filtering):
- The event exists in ’all_caption_events’ but is NOT in ‘required_events‘.
- The model mentioned it only to describe the scene or explicitly stated it was not relevant.
- Example: I hear wind in the background, but the main sound is the car engine. (Here, ’wind’ is a neutral mention, not an error).
- Action: Do NOT count these as errors.
5. missed_events (Omission):
- Events in ‘required_events‘ that are NOT found in ‘matched_events‘.

Input Data
[Question]: QUESTION
[Correct Answer]: CORRECT_ANSWER
[Ground Truth Audio Caption]:
GROUND_TRUTH_CAPTION
[Model Reasoning Path]: MODEL_REASONING

Output Format (JSON Only)
all_reasoning_events: [list of all events mentioned by model],
matched_events: [list],
error_matched: [list],
error_use: [list],
neutral_events: [list],
missed_events: [list]

B CAFE Evaluation Supplement

B.1 Test model

In Section 3, the open-source LALMs evaluated include MiMo-Audio (Xiaomi, 2025), Qwen2.5-Omni-7B (Xu et al., 2025), DeSTA2.5-Audio (Lu et al., 2025), Phi-4-Multimodal (Abouelenin et al., 2025), Qwen2-Audio-Instruct (Chu et al., 2024), as well as Omni-R1 (Rouditchenko et al., 2025) without Chain-of-Thought (CoT) RL training. Regarding LARMs, we evaluate Audio-Flamingo-3 (Goel et al., 2025), Audio Reasoner (Xie et al., 2025), Step-Audio-R1.1 and Step-Audio-R1 (Tian et al., 2025). Additionally, we incorporate the commercial models Gemini-2.5-Flash (Comanici et al., 2025) and GPT-4o-Audio (Hurst et al., 2024).

B.2 Model Inference Prompt

In CAFE evaluation experiments (Figure 3), models (c) and (d) lack CoT training, their output lengths and formats are inherently unstable. We use the "answer first, then reasoning" prompt strictly to ensure stable answer extraction. To eliminate any concerns and guarantee an absolutely fair comparison, we re-evaluated the baseline LALMs using a "reasoning first, then answer" prompt. The results demonstrate that the models still exhibit clear Audio Perception Decay. Furthermore, the overall reasoning accuracy of the baselines did not change meaningfully.

The following prompt is used to evaluate LALMs and commercial models:

[Quesiton]
Output the selected option first. Then write Reasoning: and explain your choice, mentioning the audio events that support it.
[Choices]

Prompt used to evaluate LARMs:

[Quesiton]
Select one option from the provided choices.
[Choices]

B.3 CAFE Framework Reliability Verification

Following the LLM-as-a-Judge paradigm, we employ Gemini-3-Pro for CAFE's scalable evaluation, leveraging its strong multimodal capabilities. To further assess the reliability of the automatic evaluation, we conduct human cross-validation on 100 randomly sampled instances. Specifically, human annotators verify both the caption text and the events extracted by Gemini-3-Pro, where the extracted events are evaluated across the categories of Error-matched, Error-used, Matched, Missed, and Neutral. The category-wise human–Gemini agreement is reported in Table 3 , with an average agreement rate of 92%, indicating that Gemini-3-Pro provides sufficiently reliable judgments for our evaluation setting.

To assess prompt and model stability, we varied the wording of the evaluation prompts and compared the results using GPT-5 as the underlying API. We found that the perception accuracy fluctuated by less than ±2.1%, and the relative rankings of all baseline models remained entirely consistent.

B.4 Model Experiment Detail

Our experiment classifies model responses post-hoc based on their generated length. In Figure 3, models (a) and (b) underwent CoT training to enhance their performance on high-difficulty questions. Models (c) and (d) did not receive CoT training. However, exactly like models (a) and (b), they exhibit a clear decay in audio perception accuracy as reasoning length increases. We attribute this uniformly to an insufficient perception capacity when processing complex tasks through step-by-step reasoning. This deficiency directly damages reasoning accuracy as the reasoning chain elongates.

Furthermore, we conducted a comparative experiment on the non-CoT trained models (c) and (d) without enforcing any reasoning output constraints. We found that their final accuracy remained fundamentally consistent with the CoT reasoning results, showing only a minor fluctuation of approximately ±0.8%.

	Cap.	E-mat.	E-use.	Mat.	Miss.	Neu.	Avg.
Agr.	94%	89%	91%	92%	93%	93%	92%

Table 3: Human agreement rates on 100 randomly sampled instances. The table reports agreement (Agr.) for the caption text (Cap.) and for events extracted by Gemini-3-Pro across the categories Error-matched (E-mat.), Error-used (E-use.), Matched (Mat.), Missed (Miss.), and Neutral (Neu.). The last column shows the mean agreement across all reported items.

B.5 Reasoning Prompts and Response Cases

Implicit and explicit unstructured reasoning prompts for RL:

Implicit Prompt:
Output the final answer in <ANSWER ></ANSWER>

Explicit Unstructured Reasoning Prompt:
Output the thinking process in <THINK></THINK>and final answer in <ANSWER></ANSWER>.

Response cases from different models:

Implicit Case:
Instruction:
Based on the given audio, identify the source of the whoop. Please choose the answer from the following options: ['Human', 'Bird', 'Dog', 'Machine'].
Output the final answer in <ANSWER> </ANSWER>.

Response:
<ANSWER> Human</ANSWER>

Explicit Case:
Instruction:
Based on the given audio, identify the source of the whoop. Please choose the answer from the following options: ['Human', 'Bird', 'Dog', 'Machine'].
Output the thinking process in <THINK> </THINK> and final answer in <ANSWER> </ANSWER>.

Response:
<THINK>
Hmm, this whoop sounds pretty human. It's definitely not a bird, dog, or machine. If it were a machine, it wouldn't have that kind of emotional or vocal quality. Let's see if I'm making a leap here. If it were a bird, the whoop would likely be more of a chirp or tweet. A dog might bark or whine, but certainly not whoop. So, if it's not a machine, it must be a human. That seems to fit best with what I'm hearing. Am I making a leap here? Let's see if this holds up under scrutiny. If it were a human, wouldn't the whoop be more of an emotional expression or a social interaction? Yes, that makes sense. So, it's probably a human whoop.
</THINK>
<ANSWER>Human</ANSWER>

We confirm that RL training utilizing explicit

reasoning prompts induced test-time scaling behavior. This approach significantly increased both the model's response length during the inference stage. For instance, on the MMAU benchmark, the average response length for the Qwen2.5-Omni baseline using direct answering is 11.208 tokens. In contrast, CoT-trained models built upon this same base model generate significantly longer outputs: R1-AQA produces an average reasoning length of 42.269 tokens, while the closed-source SARI model reports about 500 tokens (in its Figure 3).

C Benchmark and Framework Details

We primarily evaluate model reasoning performance using accuracy on multiple-choice questions, and further conduct reasoning analysis under the CAFE framework. The main evaluations and metrics used are:

MMAU Benchmark (Sakshi et al., 2024) We evaluate the model on the MMAU test-mini split, which features demanding audio QA tasks requiring advanced reasoning. Model performance is reported as multiple-choice accuracy. The MMAU-v05.15.25 release revises approximately 25% of the questions and answers from the original MMAU benchmark to improve clarity, accuracy, and overall quality. In addition, around 5% of the audio files are refined to enhance acoustic consistency and signal fidelity.

The official MMAU benchmark had not been updated during our initial experiments, particularly during the development of the MPAR^2 framework with different reward functions. As a result, all experiments were conducted using the origin version of MMAU. To ensure experimental consistency, all results reported in the main paper are based on this previous version.

After the release of MMAU-v05.15.25, we further evaluated the trained models on the updated benchmark to assess their performance under the latest setting and to enable direct comparison with current state-of-the-art models.

MMAR Benchmark (Ma et al., 2025b) The benchmark is designed to evaluate the reasoning capabilities of Audio Language Models in realistic acoustic environments. It consists of 1,000 audio–question–answer samples collected from open-domain videos and curated through expert annotation and quality control. Unlike prior benchmarks that focus on isolated audio categories, MMAR includes mixed and overlapping audio compositions. For fair comparison with the MMAU dataset, which reports results only on isolated categories, Tables 2 and 5 present MMAR results under isolated audio settings. Queries are further organized into four hierarchical reasoning levels, ranging from low-level acoustic analysis to high-level semantic and culturally grounded inference.

CAFE Framework As detailed in Section 3, CAFE quantifies the perception and utilization of audio events during reasoning. Leveraging ground-truth audio and QA pairs, it evaluates performance via metrics covering perception accuracy, reasoning accuracy, etc. Notably, in our event extraction design (Section 3.1), we established a specific "Neutral" category. For audio events that are present but unnecessary for answering the question (i.e., "NOT in 'required events'"), the framework explicitly instructs the evaluator: "Do NOT count these as errors". This rule applies provided the model does not incorrectly use them for logical deduction. Therefore, CAFE evaluates whether the model captures necessary information and avoids misusing irrelevant details. It does not mechanically require the model to list all voices.

D Supplementary Results

D.1 MMAU (v05.15.25) Results

As shown in Table 4, when evaluated on MMAU-v05.15.25, the MPAR^2 model achieves a substantial improvement over the baseline Qwen2.5-Omni-7B, which is consistent with the performance trends observed on the original MMAU benchmark. In addition, compared with previous models, MPAR^2 delivers highly competitive results.

D.2 Analysis of Audio-Focused Attention

We further investigate how MPAR^2 training reshapes the model's attention distribution over audio tokens. Specifically, we calculate the attention ratios between generated tokens and input audio. As illustrated in Figure 7, we visualize the attention dynamics by plotting the attention ratios of the base model and MPAR^2 across reasoning token length. The results indicate that MPAR^2 maintains a consistently higher attention level on audio compared to the vanilla Qwen2.5-Omni. Notably, attention peaks during the early reasoning stage, driven by the initial perceptual processing. Following this peak, rather than decaying, the attention exhibits wave motions, reflecting the model's continuous

Model	Sound	Music	Speech	Avg
Large Audio Language Models				
Omni-R1	81.70	73.40	76.00	77.00
MiMo-Audio	81.68	**74.25**	68.17	74.70
Qwen2.5-Omni-7B	78.10	65.90	70.60	71.50
Kimi-Audio	75.68	66.77	62.16	68.20
Qwen2-Audio-Instruct	67.27	56.29	55.26	59.60
Large Audio Reasoning Models				
Audio-Thinker	81.98	**74.25**	**76.88**	**77.70**
Step-Audio 2	**84.04**	73.56	75.15	77.58
Audio Flamingo 3	79.58	73.95	66.37	73.30
Audio-Reasoner	67.87	69.16	66.07	67.70
Commercial Models				
Gemini 2.5 Flash	73.27	65.57	76.58	71.80
GPT-4o Audio	64.56	56.29	66.67	62.50
MPAR^2-7B	79.70	73.30	75.60	76.20

Table 4: Performance Comparison on MMAU-v05.15.25 (Test-mini). Results for other methods are sourced from the MMAU Leaderboard: MMAU-v05.15.25. The best results highlighted in **bold**, and the second-best scores are underlined.

retrieval with audio context to address perception-related sub-questions. This pattern suggests that MPAR^2 enhances low-level grounding at the input stage while also reinforcing audio-conditioned reasoning in the later decoding process. Together, these results indicate that MPAR^2 encourages the model to more effectively integrate acoustic cues throughout the perception–reasoning structure.

D.3 Discussion of "Audio Perception Decay"

We argue that "Audio Perception Decay" is by no means that "long reasoning is fundamentally harmful in audio reasoning". Instead, we aim to reveal a specific bottleneck: current audio models lack the ability to stably anchor acoustic cues across long generated sequences. Perception decay prevents extended reasoning from delivering its intended positive impact. As the reasoning chain elongates, the model gradually forgets or hallucinates audio events, which degrades the accuracy of subsequent logical deductions.

Meanwhile, we posit that *forcing lengthy reasoning on simple perception tasks is unnecessary*. Our primary analysis supporting the Audio Perception Decay conclusion is based on the high-difficulty, reasoning-intensive MMAR benchmark, which features mixed overlapping audio compositions and spans four hierarchical levels from basic acoustic analysis to high-level semantic and cultural inference. It is inherently a highly difficult and reasoning-heavy benchmark.

Furthermore, on both MMAR and the perception-oriented MMAU, our MPAR^2 method

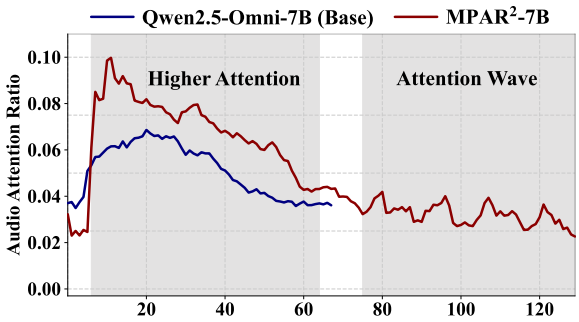


Figure 7: Audio attention ratio across reasoning token length.

significantly improves both perception and overall reasoning performance. This suggests that we do not consider long reasoning universally inefficient. Rather, by enhancing perception during reasoning, we successfully unlock the true potential of CoT in LALMs.

D.4 MPAR^2-3B Results

Performance Enhancement Trajectory Remains Consistent at the 3B Scale As detailed in Table 5, MPAR^2-3B significantly outperforms its baseline, elevating accuracy on MMAU from 63.30% to 70.17% and on MMAR from 53.80% to 55.62%. Mirroring the 7B results, widespread gains are evident across Sound, Music, and Speech domains. These findings confirm that the efficacy of MPAR^2 is agnostic to model size, effectively empowering smaller models with robust audio perception and reasoning capabilities.

Similar Mitigation of Perception Decay is also Observed at the 3B Scale As illustrated in Figure 8 (a)-(b), the MPAR^2-3B model successfully mitigates the trend of perception decay, maintaining robust accuracy even as reasoning chains expand. This consistency confirms that our framework effectively stabilizes perceptual fidelity irrespective of model size.

Adaptive Reasoning Length Persists at the 3B Scale As illustrated in Figure 8 (c), MPAR^2-3B effectively aligns its reasoning length with task complexity, generating notably longer chains for the reasoning-intensive MMAR benchmark compared to MMAU. Specifically, the model produces concise paths for simple tasks while reserving expanded trajectories for complex acoustic scenarios. This demonstrates a learned ability to dynamically scale reasoning depth, thereby optimizing the balance between computational efficiency and quality.

Model	MMAU (Test-mini)				MMAR			
	Sound↑	Music↑	Speech↑	Avg↑	Sound↑	Music↑	Speech↑	Avg↑
Qwen2.5-Omni-3B	70.27	60.48	59.16	63.30	53.94	**46.12**	53.74	53.80
Qwen2.5-Omni-7B	69.40	66.80	61.60	65.90	53.94	42.23	58.84	55.20
MPAR^2-3B	76.42	65.11	69.04	70.17	53.68	41.56	58.45	55.62
MPAR^2-7B	**77.98**	**71.30**	**74.47**	**74.59**	**62.40**	44.30	**64.26**	**60.32**

Table 5: Accuracy (%) on MMAU (original) and MMAR benchmark.

Model	Acc_{per} ↑	Err_{per} ↓	Err_{use} ↓	Err_{omit} ↓	$MMAR_{acc}$ ↑
Qwen2.5-Omni-7B	31.74	29.61	28.69	57.37	55.20
Qwen2.5-Omni-3B	28.54	39.35	30.9	52.14	53.80
MPAR^2-7B	**63.51**	**23.14**	**7.74**	**30.59**	**60.32**
MPAR^2-3B	61.04	23.39	18.74	32.46	55.62

Table 6: Results of various audio models on the MMAR benchmark within the CAFE framework. † denotes $MMAR_{acc}$ evaluated by GPT-5 judge. All results are averaged over three inference runs.

Robust Perceptual Performance in 3B Scale Table 6 reveals that this improvement is scale-invariant: MPAR^2-3B achieves a perception accuracy (Acc_{per}) of 61.04%, doubling the performance of its baseline (28.54%) and delivering results comparable to the 7B variant. This highlights the method's potential to significantly enhance perception even in smaller-scale models.

D.5 Ablation Study of Reasoning Mechanisms

Method	MMAU ↑	MMAR ↑	CAFE Acc_{per} ↑	Avg. Tokens
Base	65.90	55.20	31.74	9.21
Direct-answer GRPO	70.64	57.20	32.54	6.29
Plain CoT GRPO	68.32	54.60	27.34	549.36
MPAR^2-7B	**74.59**	**60.32**	**63.51**	497.20

Table 7: Results of direct-answer GRPO and plain CoT GRPO under the same training settings. Base model is Qwen2.5-Omni-7B. Plain CoT GRPO refers to RL training that only includes format and answer rewards. Token count statistics based on MMAR results

As shown by the ablation results in Table 7, direct-answer GRPO consistently outperforms the base, indicating that RL itself is effective for improving audio QA performance. In contrast, plain CoT RL underperforms direct-answer GRPO and yields the lowest perceptual accuracy, suggesting that simply introducing CoT reasoning is not a stable strategy for audio QA and that the problem of perception decay becomes particularly severe under this setting. Notably, MPAR^2 substantially surpasses plain CoT RL, showing that the limitation does not lie in explicit reasoning itself, but rather in the lack of perception-aware reasoning. Furthermore, MPAR^2 also outperforms direct-answer GRPO, demonstrating that explicit reasoning can be effective when guided by perceptual enhancement. Meanwhile, compared to plain CoT training, MPAR^2 reduces the token budget to some extent.

E Training Details

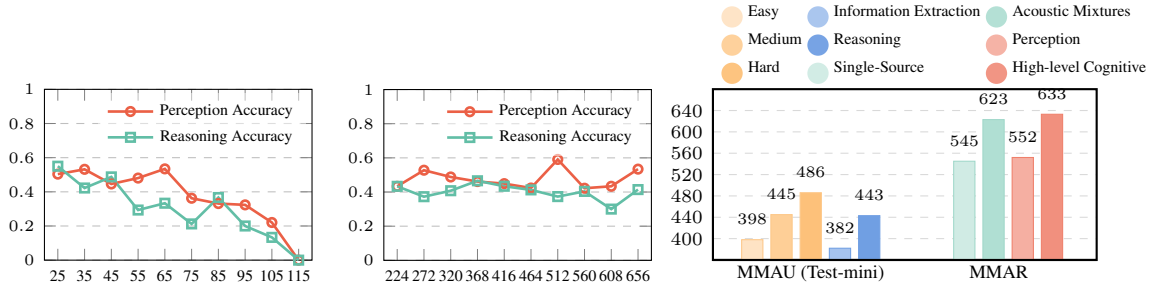
E.1 SFT and RL Setting

SFT:
Epochs: 3
Train dataset: $Data_{SFT}$
Total batch size: 16
Learning rate: 5e-5

GRPO:
Train type: LoRA
LoRA rank: 8
LoRA alpha: 32
Torch dtype: bfloat16
Max completion length: 2048
Per device train batch size: 1
Per device eval batch size: 1
Gradient accumulation steps: 2
Learning rate: 1e-6
Warmup ratio: 0.05
Num generations: 8
Temperature: 1.0
Top p: 0.99
Top k: 50
Deepspeed: ZeRO-3

E.2 Data Production Process

The overview of the data construction is in Figure 9. To teach LALMs this structured reasoning pattern and prepare data for RL, we draw from the AVQA dataset (Yang et al., 2022), which is designed for video question answering. We first employed Gemini-2.5-Pro to generate detailed captions for the AVQA dataset. Based on these captions, Qwen3-32B synthesized reasoning-oriented QA pairs across four dimensions: timbre characteristics, temporal patterns, pitch contours, and rhythmic structures. Then, we filter the QA pairs for logic and consistency, and assess difficulty by running the initial base model 16 times, discarding



(a) Qwen2.5-Omni-3B (Base) (b) MPAR^2-3B (c) Reasoning Token Length of MPAR^2-3B on Benchmarks

Figure 8: (a)-(b) is the accuracy of MPAR^2-3B vs. Base across token lengths. (c) is the Reasoning token length of MPAR^2-3B across different task categories on the MMAU and MMAR benchmark.

samples that were completely correct or incorrect, resulting in 46,544 QA samples. From the filtered data pool, we curated a balanced set of 5,000 samples across four aspects and different time durations, while the remaining 41,544 samples $Data_{RL}$ were reserved for Stage-2 RL. Finally, we prompt to use these 5,000 samples to generate CoT reasoning paths by Qwen3-32B. Following CoT quality filtering, 4,600 instances, denoted as $Data_{SFT}$, were selected for Stage-1 supervised fine-tuning.

Below are the prompts used in the production process.

E.2.1 Gemini-2.5-Pro Caption Prompt

You are an expert audio event analyst. Your task is to carefully analyze the entire audio and extract all meaningful events as can as possible, including source entities, time intervals and captions.
For each extracted event, you must:
1. Identify the possible sound sources within the event and make a brief descriptions for it.
2. Provide the accurate time interval [start time,end time] for each event when it occurs within the audio.
3. Provide a detailed, precise, and contextually rich caption of what happens in that event interval.
Do not infer or imagine information that cannot be clearly heard. The max number of extracted events is 5.

Please separate different parts using the following format:
<event1>
<description>...</description>
<time>[start time, end time]</time>
<caption>...</caption>
</event1>
<event2>
<description>...</description>
<time>[start time, end time]</time>
<caption>...</caption>
</event2>
<event3>
<description>...</description>
<time>[start time, end time]</time>
<caption>...</caption>
</event3>
.....

Here is a sample. Please strictly follow the format in the sample.
<event1>
<description>dogs barking </description>
<time>[0.00s, 2.35s]</time>
<caption>the audio captures a brief, intense episode of multiple dogs barking aggressively in an outdoor setting.
</caption>
</event1>
<event2>
<description>...</description>
<time>[start time, end time]</time>
<caption>...</caption>
</event2>
<event3>
<description>...</description>
<time>[start time, end time]</time>
<caption>...</caption>
</event3>.....

Please extract events from the follow audio:

E.2.2 QA Generate Prompt

Event Counting QA Generate Prompt:

We are constructing training data to enhance audio perception and reasoning in audio large language models (LLMs). Your job is to create a single, high-quality multiple-choice question that tests whether a model can perform numerical reasoning and quantitative analysis by **listening** to a complete audio clip.

Input Audio Events (Chronological Order): events_description

Task Description:
You are provided with the ground truth sequence of sound events above. Your goal is to generate ONE multiple-choice question (MCQ) focusing on **Event Counting** & **Numerical Reasoning**.

Step 1: Suitability Check (CRITICAL)
First, analyze the event list provided above. Ask yourself:
1. Are there distinct, countable discrete events (e.g., 'bark', 'gunshot', 'footstep') rather than just continuous ambience (e.g., 'wind', 'static', 'silence')?
2. Is there enough variety or repetition to form a valid numerical question (e.g., counting total occurrences, comparing counts of two different sources)?

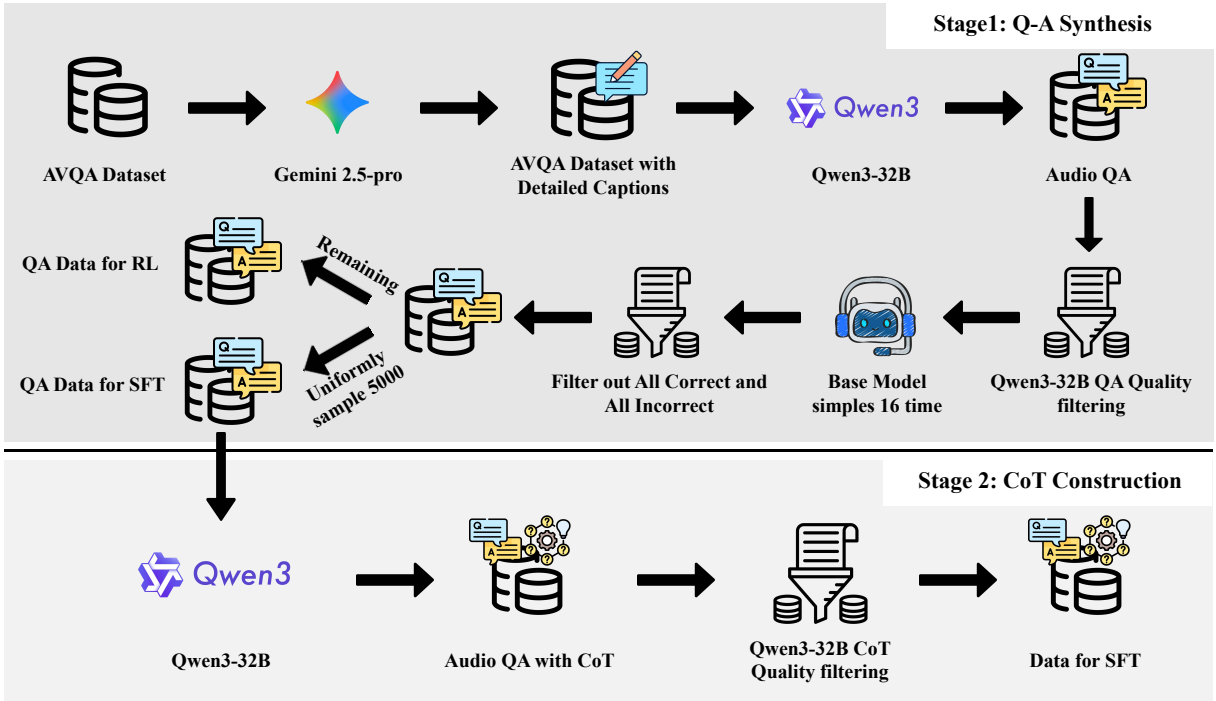


Figure 9: The overview of the data construction pipline.

If the content consists mainly of continuous noise, ambiguous sounds, or a single non-repeatable event that makes counting trivial or impossible, **you must output exactly**: Not suitable for this hallucination type.

Step 2: Question Generation (If Suitable)
If the audio events support numerical reasoning, generate ONE MCQ.
The question must target the **auditory experience**. Do NOT ask about the text descriptions directly. The question must be answerable by listening to the audio and counting/analyzing the sounds.

Focus Areas:
- **Total Count**: 'How many times is the [specific sound] heard?'
- **Source Comparison**: 'Did the dog bark more times than the cat meowed?'
- **Sequence Logic**: 'After the first door slam, how many footsteps follow?'
(Note: Since you do not have exact timestamps, focus on the count and order of events, not the speed or rate per minute.)

STRICT CONSTRAINT:
The question must NOT contain phrases like 'According to the list', 'In the description', or 'text'. It must sound like a natural question asked to someone who has just listened to the recording.

Output Format (if suitable):
Question: <question about the sound itself>
A. <option A>
B. <option B>
C. <option C>
D. <option D>
Correct answer: <the correct option letter>

Case Study of Event Counting

Question: How many distinct cawing events are clearly heard before the audio becomes chaotic with overlapping caws?
Choice: A. One B. Two C. Three D. Four
Correct answer: C

Pitch Contour QA Generate Prompt

We are constructing training data to enhance audio perception and reasoning in audio large language models (LLMs). Your job is to create a single, high-quality multiple-choice question that tests whether a model can perform deep inference based on the **dynamic pitch contour (frequency modulation)** by **listening** to a complete audio clip.

Input Audio Events (Chronological Order): events_description

Task Description:
You are provided with the ground truth sequence of sound events above.
Your goal is to generate ONE multiple-choice question (MCQ) focusing on **Pitch Contour & Semantic Inference**.

Step 1: Suitability Check (CRITICAL)
Analyze the event list provided above. To support a question about pitch contour, the audio must contain sources with detectable tonal properties or modulation. Ask yourself:
1. Does the audio contain **speech** (where intonation conveys meaning like sarcasm or questions)?
2. Does it contain **tonal machinery or vehicles** (where speed/movement affects pitch, e.g., Doppler effect, acceleration)?
3. Does it contain **musical or biological sounds** (e.g., bird calls, singing) where frequency changes involve meaning?

If the audio consists ONLY of unpitched sounds (e.g., 'rain', 'static noise', 'footsteps', 'door slam', 'wind') where pitch analysis is irrelevant or impossible, **you must output exactly:**
Not suitable for this hallucination type.

Step 2: Question Generation (If Suitable)
If the audio contains tonal or pitch-varying events, generate ONE MCQ.
The question must require the model to hear the curve of the sound (rising, falling, wavering) and deduce the underlying cause, intent, or movement.

Focus Areas (Audio-Centric):
- **Intonation & Intent:** 'The speaker's pitch rises sharply at the end. What does this suggest about their certainty?' (Reasoning: Statement vs. Question)
- **Doppler Effect (Physics):** 'As the [vehicle] sound gets louder and then fades, the pitch drops significantly. What does this confirm about its movement?' (Reasoning: Passing the listener)
- **Mechanical State:** 'The pitch of the engine continuously increases without dropping. What does this imply about the machine's operation?' (Reasoning: Acceleration/Revving)
- **Emotional State:** 'The tremor or wavering pitch in the voice suggests what emotion?' (Reasoning: Fear/Excitement

STRICT CONSTRAINT:
The question must NOT contain phrases like 'According to the list', 'In the description', or 'text'.
It must sound like a natural question asked to someone who has just closed their eyes and listened to the recording.

Output Format (if suitable):
Question: <question about the pitch/tone/intonation>
A. <option A>
B. <option B>
C. <option C>
D. <option D>
Correct answer: <the correct option letter>

Case Study of Pitch Contour

Question: A siren is heard starting low, rising quickly to a high pitch, and then holding that pitch for a while before stopping. What is this pattern most likely indicating?
Choice: A. A fire truck arriving at the scene B. An emergency evacuation alert C. A vehicle accelerating away D. A malfunctioning car alarm
Correct answer: B

Rhythmic Structure QA Generate Prompt

We are constructing training data to enhance audio perception and reasoning in audio large language models (LLMs). Your job is to create a single, high-quality multiple-choice question that tests whether a model can perform deep inference based on the **rhythmic structure and temporal regularity** by **listening** to a complete audio clip.

Input Audio Events (Chronological Order): events_description

Task Description:
You are provided with the ground truth sequence of sound events above. Your goal is to generate ONE multiple-choice question (MCQ) focusing on **Rhythmic Structure & Behavioral Inference**.

Step 1: Suitability Check (CRITICAL)
Analyze the event list provided above. To support a question about rhythm, the audio must contain sounds that repeat or form a pattern. Ask yourself:
1. Does the audio contain **repetitive impulsive sounds** (e.g., 'footsteps', 'typing', 'heartbeat', 'clapping', 'knocking')?
2. Does it contain **rhythmic machinery or engines** (e.g., 'train wheels', 'clock ticking', 'idling engine')?
3. Does it contain **music or percussion**?
If the audio consists ONLY of continuous amorphous noise (e.g., 'wind', 'water flow', 'static') **OR single isolated events** (e.g., 'one gunshot', 'a single drop', 'one scream') where no rhythm exists, **you must output exactly:**
Not suitable for this hallucination type.

Step 2: Question Generation (If Suitable)
If the audio contains rhythmic patterns, generate ONE MCQ.
The question must require the model to hear the pattern of the sound (regularity, tempo, acceleration, chaos) and deduce the underlying behavior or state.

Focus Areas (Audio-Centric):
- **Activity State (Tempo):** 'The footsteps transition from a slow, steady walking pace to a rapid, frantic rhythm. What does this suggest about the subject?' (Reasoning: Leisure vs. Fleeing)
- **Stability (Regularity):** 'The engine sound changes from a steady hum to an irregular, sputtering pattern. What does this indicate?' (Reasoning: Mechanical failure)
- **Coordination (Synchronization):** 'The clapping sounds represent a large group hitting the beat in perfect unison. What context does this imply?' (Reasoning: Organized audience/Performance vs. Chaotic crowd)
- **Environmental Context:** 'The rhythm of the typing is sporadic and hesitant, with long pauses. What does this suggest about the typist?' (Reasoning: Thinking/Uncertainty vs. Professional transcription)

STRICT CONSTRAINT:
The question must NOT contain phrases like 'According to the list', 'In the description', or 'text'. It must sound like a natural question asked to someone who has just closed their eyes and listened to the recording.

Output Format (if suitable):
Question: <question about the pitch/tone/intonation>
A. <option A>
B. <option B>
C. <option C>
D. <option D>
Correct answer: <the correct option letter>

Case Study of Rhythmic Structure

Question: The pig's vocalizations occur in a tightly spaced, repetitive sequence at the start of the audio. What does this rhythmic pattern most likely suggest about the pig's emotional or behavioral state? A. It is calmly and slowly exploring its environment.
Choice: A. It is calmly and slowly exploring its environment. B. It is experiencing distress or agitation. C. It is communicating with a distant group of pigs. D. It is playing or engaging in a harmless interaction with another

animal.
Correct answer: B

Temporal Event Reasoning QA Generate Prompt

We are constructing training data to enhance audio perception and reasoning in audio large language models (LLMs). Your job is to create a single, high-quality multiple-choice question that tests whether a model can perform **Temporal Logic and Causal Reasoning** by **listening** to a complete audio clip.

Input Audio Events (Chronological Order): events_description

Task Description:
You are provided with the ground truth sequence of sound events above. Your goal is to generate ONE multiple-choice question (MCQ) focusing on **Sequence, Causality, or Temporal Relationship**.

Step 1: Suitability Check (CRITICAL)
Analyze the event list provided above. To support temporal or causal reasoning, the audio must contain a **sequence of multiple distinct events**. Ask yourself:
1. Are there **at least two distinct events** happening in succession (e.g., 'thunder' then 'rain', 'footsteps' then 'door open')?
2. Is there a logical link (cause-and-effect) or a clear chronological order to test?
If the audio consists of a SINGLE event (e.g., just 'dog barking') OR only continuous background noise (e.g., 'static', 'city traffic') where no sequence exists, **you must output exactly:**
Not suitable for this hallucination type.

Step 2: Question Generation (If Suitable)
If the audio contains a valid sequence of events, generate ONE MCQ.
The question must require the model to **track the order of sounds** or **deduce the cause** of a sound based on what happened before/after.
(Note: Do NOT ask about specific duration in seconds, as you do not have exact timestamps. Focus on relative order and logic.)

Focus Areas (Audio-Centric):
- **Chronological Order:** 'Which sound occurred immediately after the glass shattered?' (Reasoning: Tracking sequence).
- **Causality/Reaction:** 'The sudden braking sound was immediately followed by a crash and shouting. What does this sequence imply?' (Reasoning: Cause and Effect).
- **Interruption:** 'The music was abruptly cut off by which specific sound?' (Reasoning: Identifying the disruptor).
- **Reverse Inference:** 'The sound of the audience clapping suggests that what event likely just finished?' (Reasoning: Inferring the preceding context).

STRICT CONSTRAINT:
The question must NOT contain phrases like 'According to the list', 'In the description', or 'text'. It must sound like a natural question asked to someone who has just closed their eyes and listened to the recording.

Output Format (if suitable):
Question: <question about the pitch/tone/intonation>
A. <option A>
B. <option B>
C. <option C>
D. <option D>
Correct answer: <the correct option letter>

Case Study of Etemporal logic

Question: A single dog begins barking, and soon after, multiple dogs join in. Which of the following sounds is most likely to have triggered the initial barking?
Choice: A. The low-frequency electronic tone B. The person sighing or groaning C. The sound of footsteps D. The rustling of clothing
Correct answer: A

Timbre Characteristics QA Generate Prompt

We are constructing training data to enhance audio perception and reasoning in audio large language models (LLMs). Your job is to create a single, high-quality multiple-choice question that tests whether a model can perform **Temporal Logic and Causal Reasoning** by **listening** to a complete audio clip.

Input Audio Events (Chronological Order): events_description

Task Description:
You are provided with the ground truth sequence of sound events above. Your goal is to generate ONE multiple-choice question (MCQ) focusing on **Sequence, Causality, or Temporal Relationship**.

Step 1: Suitability Check (CRITICAL)
Analyze the event list provided above. To support temporal or causal reasoning, the audio must contain a **sequence of multiple distinct events**. Ask yourself:
1. Are there **at least two distinct events** happening in succession (e.g., 'thunder' then 'rain', 'footsteps' then 'door open')?
2. Is there a logical link (cause-and-effect) or a clear chronological order to test?
If the audio consists of a SINGLE event (e.g., just 'dog barking') **OR only continuous background noise** (e.g., 'static', 'city traffic') where no sequence exists, **you must output exactly:**
Not suitable for this hallucination type.

Step 2: Question Generation (If Suitable)
If the audio contains a valid sequence of events, generate ONE MCQ.
The question must require the model to **track the order of sounds** or **deduce the cause** of a sound based on what happened before/after.
(Note: Do NOT ask about specific duration in seconds, as you do not have exact timestamps. Focus on relative order and logic.)

Focus Areas (Audio-Centric):
- **Chronological Order:** 'Which sound occurred immediately after the glass shattered?' (Reasoning: Tracking sequence).
- **Causality/Reaction:** 'The sudden braking sound was immediately followed by a crash and shouting. What does this sequence imply?' (Reasoning: Cause and Effect).
- **Interruption:** 'The music was abruptly cut off by which specific sound?' (Reasoning: Identifying the disruptor).
- **Reverse Inference:** 'The sound of the audience clapping

suggests that what event likely just finished?' (Reasoning: Inferring the preceding context).

STRICT CONSTRAINT:
The question must NOT contain phrases like 'According to the list', 'In the description', or 'text'.
It must sound like a natural question asked to someone who has just closed their eyes and listened to the recording.

Output Format (if suitable):
Question: <question about the pitch/tone/intonation>
A. <option A>
B. <option B>
C. <option C>
D. <option D>
Correct answer: <the correct option letter>

Case Study of Timbre Characteristics

Question: Based on the sound characteristics of the buzzing in the background, what is the most likely physical source producing this continuous, low-frequency noise?
Choice: A. A swarm of bees inside a wooden beehive B. A malfunctioning electrical device emitting a hum C. A distant, low-flying airplane D. A person humming a tune consistently and softly
Correct answer: A

E.3 QA Fliter Prompt

Role
You are an expert Data Quality Evaluator for Audio-Text Reasoning Datasets. Your goal is to strictly filter out low-quality QA pairs and keep only those that require genuine logical reasoning.

Objective
Evaluate the following Audio QA pair. You must determine if the Question requires logical deduction, causal inference, or temporal analysis of the events described in the Caption.

Input Data
- **Audio Caption**: caption
- **Question and Answer**: question

Evaluation Criteria (The ReasoningLitmus Test)

1. POSITIVE INDICATORS (High Score / KEEP)
The question requires the model to:
Infer Cause/Effect: Why did the sound stop?(Requires understanding the preceding event).
Analyze Sequence (Temporal): What happened immediately after the explosion?
Deduce State/Intent: Based on the footsteps and breathing, is the person running or walking?
Synthesize Multiple Clues: Combining background noise + specific actions to determine the location.

2. NEGATIVE INDICATORS (Low Score / DISCARD)
Simple Pattern Matching: The answer is just a word lifted directly from the caption (e.g., Caption says a red car, Question asks what color is the car?).
Common Sense / General Knowledge: Can be answered without the audio caption (e.g., Do birds fly?).
Summarization: Describe the audio(This is generation, not reasoning).
Unsolvable/Hallucination: The answer assumes facts not present in the caption.

Task
1. Analyze the relationship between the Caption, Question, and Answer.
2. Assign a **Score (1-5)** based on reasoning depth.
- **5**: Complex reasoning (multi-hop, causal, or temporal).
- **4**: Clear deduction required.
- **3**: Simple inference.
- **2**: Direct text retrieval / keyword matching.
- **1**: Common sense or irrelevant.
3. Make a final **Decision**: KEEP(Score >= 4) or DISCARD(Score <4).

Output Format
Return ONLY a valid JSON object:
analysis: Brief explanation of the reasoning logic required.,
score: <int>,
decision: KEEPor DISCARD

E.4 MPAR^{2}CoT Generate Prompt

Role:
You are an advanced AI assistant specializing in Audio Reasoning and Chain-of-Thought (CoT) synthesis. You also act as a strict auditor to ensure data quality.

Task:
Your task is to synthesize the provided input data into a structured CoT format using specific XML tags, and then perform a **critical validation (Review)** of the reasoning process.
You will be given:
1.**Question:** The main query about the audio.
2.**Final Answer:** The ground truth answer
3.**Caption:** A time-interval detailed description of the audio events (Ground Truth)
4.**Model Outputs:** A list of step-by-step sub-questions and answers generated logically in previous steps.

Output Format Requirements:
You must output the content strictly inside the following XML structure:
<thinking>
<perception>
1. [start_time, end_time]: Description of event A.
2. [start_time, end_time]: Description of event B.
... (List ALL events from the Caption, chronologically)
</perception>
<reasoning>
1. Sub-question: [First step from Model Outputs]
Answer: [Answer to first step]
... (Include all steps from the Model Outputs)
</reasoning>
<review>
1. Evidence Check: [Simulate a **Re-listening**process. Verify if the audio events cited in the 'Reasoning' are factually supported by the events listed in 'Perception'. Check for hallucinations, misinterpretations, or missing details. **Note: Treat the 'Perception' content as the audio itself; do not refer to 'captions', 'text', or 'provided descriptions'.**]
2. Logic Check: [Evaluate the logical validity. Does the conclusion naturally follow from the evidence? Is the overall chain coherent?]
</review>
</thinking>
<answer>
[The Final Answer provided in the input]
</answer>

Directives:
1.**Perception Section (Full Extraction):** Translate the **ENTIRE** provided 'Caption' into a structured, chronological list of time intervals and event descriptions. **Do not filter, summarize, or omit** any events, even if they seem irrelevant to the specific Question. Format: '[t1, t2]: content'.
2.**Reasoning Section:** Directly utilize the provided 'Model Outputs'. Organize them into a numbered list of Subquestionand Answerpairs. Ensure the reasoning logic flows smoothly.
3.**Review Section (Critical Audit):**
Evidence Check: Perform a **Re-Perception** check.
Context: Imagine you are re-checking the audio stream directly.
Validity: Did the reasoning cite sounds that actually exist in the 'Perception' list?
Accuracy: Did the reasoning interpret the sound properties correctly?
Constraint: Strictly avoid phrases like according to the captionor the text says. Instead, use phrases like The audio contains..., I hear..., or The event at [timestamp] shows....
Logic Check: Verify the soundness of the deductive process. Ensure there are no logical leaps or contradictions.
4.**Answer Section:** State the final answer clearly.
Input Data:
Question:
ORIGINAL_QUESTION
Final Answer:
FINAL_ANSWER
Caption:
caption_w_time
Model Outputs (Reasoning Chain):
sub_question_list_generated

E.5 CoT Data Fliter Prompt

Role
You are an expert evaluator for Audio-Language Models. Your task is to audit a Chain-of-Thought(CoT) process generated by an AI model. You will assess how well the model hearsthe audio (Perception), thinksabout it (Reasoning), and auditsits own conclusion (Review).
Input Data
1. **Original Question**: The user's query.
2. **Final Answer**: The ground truth answer.
3. **Caption**: Ground truth events with precise time ranges.
4. **CoT**: The step-by-step reasoning process containing <perception>, <reasoning>, and <review>tags.

Evaluation Dimensions
Dimension 1: Perception Evaluation (The Ear)
Analyze specific claims about audio events/timestamps in the <perception>and <reasoning>sections.
Evaluate based on:
1.**Accuracy**: Does the event described match the Caption?
2.**Hallucination/Omission**: Are there invented sounds or missed critical sounds?
Dimension 2: Reasoning Evaluation (The Brain)
Analyze the logical flow in the <reasoning>section.
Evaluate each step on:
1.**Utility**: Is this step necessary for solving the Original Question?
2.**Factuality**: Is the statement factually true based on the audio content?
3.**Logical Validity**: Does the conclusion naturally follow from the cited evidence?
Dimension 3: Review Evaluation (The Auditor)
Analyze the <review>section, specifically the Evidence Checkand Logic Check.
Evaluate based on:
1.**Evidence Re-verification**:
Did the model correctly re-examine and re-perceive the audio events cited in the reasoning?
Did it accurately confirm whether the events exist in the perception data?
Critical: Did it successfully identify valid evidence versus hallucinated evidence?
2.**Rationality Check**: Did the model correctly assess the logical coherence of the entire chain? Did it ensure that the Final Answer is the only logical conclusion derived from the evidence?

Scoring & Output Format
Output only two numbers (0-10) strictly in accordance with the following format: Reasoning_Score/Review_Score
Scoring Criteria:
Reasoning_Score (num1): Rate the quality of Dimension 1 (Perception) and Dimension 2 (Reasoning). 10 = Perfect audio detection and flawless logic.
Review_Score (num2): Rate the quality of Dimension 3 (Review). 10 = The model performed a rigorous, accurate self-audit that correctly validated the evidence and logic. 0 = The review was superficial, inaccurate, or failed to catch obvious errors.

Input Data:

Question: ORIGINAL_QUESTION

Final Answer: FINAL_ANSWER

Caption: caption_w_time

Model Outputs (Reasoning Chain):
sub_question_list_generated

F Reward prompt

F.1 Perception Score Prompt

You are an expert audio perception evaluator. I will give you a record containing:
1. A Detailed Audio Caption (Ground Truth): A comprehensive, factual text description of the audio events.
2. A Question and its Correct Answer: To determine which audio events are 'critical' for the task.
3. A Model Perception Output: The content within the `<perception>` tags generated by the model, describing events with timestamps.
Your task is to evaluate the fidelity, precision, and completeness of the Model Perception against the Ground Truth, and output a single numeric score from 0, 0.1, 0.2, ..., 1.0. You must output only the score with no explanation or extra text.

Evaluate based on the following CRITICAL principles:
1. Audio Hallucination (Strict) — The model must NOT report events that do not exist in the 'Detailed Audio Caption'.
Reporting a sound that is completely absent (e.g., hearing a siren when the description only mentions birds) is a fatal failure.
2. Content Accuracy & Sequential Logic — While evaluating the model's generated timestamps, focus on:
- Event Identity: Does the model correctly identify the sound sources described in the Ground Truth? (e.g., distin-

guishing 'footsteps' from 'knocking').
- Chronological Flow: Does the sequence of events in the model's output match the narrative order of the Ground Truth? (e.g., if the description says 'a door opens then slams', the model must not place the slam before the opening).
3. Critical Event Coverage (Relevance) — The model must capture all 'Key Events' necessary to answer the provided 'Question'. Compare with the Question/Answer pairs: if the answer depends on a specific sound cue, omitting this specific event in the <perception>phase is a critical failure.
4. Consistency & Identity — The model should describe the same audio source consistently across different timestamps (unless the sound evolves). Avoid contradictory descriptions for the same ongoing event.
5. Redundancy & Conciseness — The perception output should be dense and informative. Penalize distinct 'loops' (repeating the exact same phrase for adjacent timestamps) or extreme verbosity that adds no new details.

Scoring guideline
1.0 = Flawless. Perfectly matches the Ground Truth description. Events are correctly identified and listed in the correct logical order. Concise.
0.8-0.9 = Excellent. Accurate detection of all key events described. The sequence is logical. Maybe minor verbosity.
0.5-0.7 = Mediocre. The KEY event was detected, but the description is vague, or information irrelevant to the question has been omitted.
0.2-0.4 = Poor. Misses a KEY event needed for the Answer, or misidentifies a sound source. Sequence is disorderly compared to the description.
0.0-0.1 = Severe. HALLUCINATION (inventing sounds not in the description), or total failure to identify the main audio event.

Penalty guideline (Apply these cumulatively to reduce the score):
[CRITICAL PENALTY] (Set Score to 0.0 $\sim$ 0.2) :
- Hallucinating an event not present in the Ground Truth description.
- Misidentifying the main sound source (e.g., 'gunshot' vs 'drum').
- Missing the specific audio cue required to answer the Question.
[MODERATE PENALTY] (-0.3 $\sim$ -0.5) :
- Sequential Logic Error (Events are listed in an order contradicting the description).
- Significant omission of details mentioned in the description.
[MINOR PENALTY] (-0.1 $\sim$ -0.2) :
- Excessive wordiness or repetitive phrasing without new information.
- Vague descriptions (e.g., 'noise' instead of 'dog barking') if the Ground Truth is specific.

Operational rule: Always output only one score (0-1 in 0.1 increments). Now evaluate the following record and output only the score.
The Detailed Audio Caption (Ground Truth) is:caption_text
The Question is:question_text
The Correct Answer is:answer_text
The Model perception to evaluate is:cot_text

F.2 Step-Level Reasoning Score Prompt

You are an expert logic and reasoning evaluator for Audio-LLMs. I will give you a record containing:
1. A Detailed Audio Caption (Model Perception): A comprehensive text description of the audio events.
2. A User Question and Context: The goal of the reasoning.
3. A Reasoning History: The steps taken so far.
4. The CURRENT STEP: The specific sub-question or reasoning step to evaluate now.

Your task is to evaluate the **validity, necessity, and audio-grounding** of the CURRENT STEP only, and output a single numeric score from 0, 0.1, 0.2, ..., 1.0. You must output only the score with no explanation or extra text
Evaluate based on the following Micro-Level dimensions:

— CRITERIA: Local Quality Check —
1. Usefulness: Is this specific reasoning step useful for answering the main question?
2. Evidence-Based Conclusion: Is the content of this step supported by the provided 'Audio Caption'?
- Every claim must align with specific events, sound sources, or acoustic details described in the caption.
3. Criticality & Efficiency: Is this step a logical next move based on the [Reasoning History]?
- Penalize 'tangential reasoning' (analyzing irrelevant noise) or redundant repetition of previous steps.

Scoring guideline:
1.0 = Perfect. The step is firmly grounded in the caption, necessary, and logically follows the history.
0.8-0.9 = Strong. Good step, but maybe slightly inefficient or the evidence citation is slightly vague.
0.5-0.7 = Mediocre. Relevant, but weak grounding (making assumptions not explicitly in the caption). Logic holds but is messy.
0.2-0.4 = Weak. The step makes a claim not supported by the caption, or merely repeats previous steps without adding value.
0.0-0.1 = Failed. Completely incoherent, visual hallucination, or factual contradiction with the caption (e.g., claiming a sound exists when caption implies silence).

Penalty guideline (Apply these cumulatively to reduce score):
[CRITICAL PENALTY] (Set Score to 0.0 $\sim$ 0.1):
- 'Factual Contradiction': The step claims a specific sound or event occurs which is explicitly absent or contradicted by the Caption.

Operational rule: Always output only one score (0-1). If [Current Step] is empty, return 0.0.

The Detailed Audio Caption is:caption_text
The User Question is:question_text
The Reasoning History is:history_text
The CURRENT STEP to evaluate is:current_step_text

F.3 Holistic Level Reasoning Score Prompt

You are an expert logic and reasoning evaluator for Audio-LLMs. I will give you a record containing:
1. A Detailed Audio Caption (Model Perception): A comprehensive text description of the audio events.
2. A Question and its Correct Answer.
3. The COMPLETE Model Reasoning: The entire chain of thought generated by the model.

Your task is to evaluate the **logical architecture, coherence, efficiency, and final derivability** of the entire process, and output a single numeric score from 0, 0.1, 0.2, ..., 1.0. You must output only the score with no explanation or extra text.

Evaluate based on the following Macro-Level dimensions:

— CRITERIA: Holistic Logical Architecture —
1. Goal-Orientation: Is the reasoning path linear and directed towards the [Correct Answer]?
- Penalize circular logic.
2. Causal Dependency: Does Step B legitimately follow Step A?
- Penalize 'Logic Jumps' where a conclusion appears out of nowhere without a preceding premise defined in the audio caption.
3. Error Propagation Check: Does an early error render the rest of the chain invalid?
- If Step 1 is wrong (e.g., misidentifying a gender or sound source compared to the caption), and subsequent steps rely on it, the whole chain collapses.
4. Final Derivability: Does the reasoning naturally flow to the [Correct Answer]?
- The conclusion must be the inevitable result of the reasoning steps, not a sudden guess.
5. Efficiency & Conciseness: Is the length of the reasoning proportional to the complexity of the question?
- **Penalize 'Over-Analysis'**: If the question is simple (e.g., 'Is there a dog?'), the reasoning should be short. Writing a 500-word essay for a simple question is a failure.
- **Penalize Repetition**: Check if the model repeats the same analysis in different words just to make the chain longer.

Scoring guideline:
1.0 = Perfect. Every step is necessary, concise, and the logic flows flawlessly from the caption evidence to the correct conclusion.
0.8-0.9 = Strong. Good logic, but maybe slightly verbose or includes one unnecessary step, yet the path is valid.
0.5-0.7 = Mediocre. Logic holds but is **bloated** or unfocused. Contains repetitive analysis or over-explains simple facts found in the caption.
0.2-0.4 = Weak. Major Logic Jumps, or the reasoning is **excessively long** and tedious without adding value (Filibustering).
0.0-0.1 = Failed. The reasoning contradicts the final answer, relies on 'Fatal Error Propagation', or is complete nonsense.

Penalty guideline (Apply these cumulatively):
[CRITICAL PENALTY] (Set Score to 0.0 $\sim$ 0.2):
- 'Fatal Error Propagation': Early false premise (contradicting the Audio Caption) corrupts the entire remaining chain.
- 'Contradiction': The reasoning concludes something different from the actual Correct Answer provided.
[MODERATE PENALTY] (-0.2 $\sim$ -0.4) :
- 'Bloated Reasoning': The reasoning is too long for the problem's difficulty (e.g., 10 steps for a Yes/No question).
- 'Irrelevance': Wasting steps on analyzing audio events that are present in the caption but do not help answer the specific Question.

Operational rule: Always output only one score (0-1).

The Detailed Audio Caption is:caption_text
The Question is:question_text
The Correct Answer is:answer_text
The COMPLETE Model Reasoning is:full_reasoning

F.4 Review Score Prompt

You are a critical meta-evaluator for the "Self-Correction" (Review) phase of an Audio-LLM.
Your task is to judge whether the [Review Content] effectively audits, verifies, and corrects the [Model Reasoning].

Input Data:
1. [Detailed Audio Caption] (Model Perception): caption_text
2. [Ground Truth Annotations] (Fact Reference): ground_truth_text
3. [User Question]: question_text
4. [Correct Answer] (Ground Truth): answer_text
5. [Model Reasoning] (Target to Audit): reasoning_text
6. [Review Content] (The Audit Output): review_text
Task: Output a single score 0.0, 0.1, ..., 1.0 for the [Review Content].

Evaluation Criteria (Review Quality):
1. Evidence Verification (Content Alignment):
- Does the Review explicitly verify that every event cited in the [Model Reasoning] actually exists in the [Detailed Audio Caption]?
- Did it catch "Hallucinations" where the Reasoning cites a sound (e.g., "dog barking") that is completely absent from the Caption?
- **CRITERIA**: The Review must confirm that the **evidence** used in reasoning is physically present in the text description.
2. Temporal & Causal Logic Audit:
- Does the Review check the narrative sequence? (e.g., if Reasoning says "A causes B", did the Review check if the Caption describes A happening before or leading into B?)
- Did it catch chronological errors where the Reasoning flips the order of events described in the text?
3. Logical Integrity Check: - Does the Review ensure the conclusion is strictly derived from the perceived evidence? - Did it flag any "Logic Jumps" (conclusions without premises) in the Reasoning?
4. Genuine Error Correction & Answer Verification (Anti-Rubber-Stamping): - **ANSWER CHECK (CRITICAL)**: Compare the conclusion/answer in [Model Reasoning] with the [Correct Answer].
- If Reasoning leads to a WRONG answer, the Review **MUST** detect and flag this failure.
- If Reasoning leads to a CORRECT answer, the Review must validate the logic flow.
- **HALLUCINATION CHECK**: If [Model Reasoning] contains factual errors (contradictions with Caption), the Review **MUST** point them out.
- **PENALTY RULES**: - **Score 0.0 IMMEDIATELY** if the [Model Reasoning] concludes with a wrong answer (mismatch with [Correct Answer]) but the Review states "The reasoning is correct" or "The answer is valid". - **Score 0.0 IMMEDIATELY** if the Review approves reasoning that contradicts the Audio Caption (e.g., approving a sound that isn't in the description).

Scoring Guide:
- 1.0: Perfect Audit. The review rigorously checked evidence presence, sequential logic, AND accurately validated the final answer against Ground Truth.
- 0.8-0.9: Strong. Good check, caught major issues, but maybe missed a minor detail in the description.

- 0.5-0.7: Generic Validation. Says "Logic is good" without citing specific text evidence. (Rubber-stamping).
- 0.2-0.4: Weak. Fails to catch obvious hallucinations or logic jumps.
- 0.0-0.1: [FATAL] The Review acts as a "Yes-Man" (Rubber-Stamp) for an INCORRECT Answer. It approves a reasoning path that leads to a result different from the [Correct Answer].
Output Rule: Return ONLY the numeric score.

G AI Tools

We used GPT-5.5 solely for writing refinement. All research ideas, experimental design, experiments, and analyses were conducted by the authors.