Listen, Pause, and Reason: Toward Perception-Grounded Hybrid Reasoning for Audio Understanding

Anonymous ACL submission

Abstract

Recent Large Audio Language Models have demonstrated impressive capabilities in audio understanding. However, they frequently suffer from perceptual errors, while reliable audio reasoning is unattainable without first grounding the model's perception in structured auditory scenes. Inspired by Auditory Scene Analysis, we first introduce PAQA, a dataset for Perception-Aware Question Answering. PAQA implements a hierarchical decoupling strategy that separates speech from environmental sound and distinguishes multiple speakers, providing explicit perceptual reasoning for training. Building on this, we propose HyPeR, a two-stage Hybrid Perception-Reasoning framework. In Stage I, we finetune the model on PAQA to perceive acoustic attributes in complex audio. In Stage II, we leverage Group Relative Policy Optimization to refine the model's internal deliberation. We introduce PAUSE tokens to facilitate latent computation during acoustically ambiguous phases and design Perceptual Consistency Reward to align reasoning rationales with raw audio. Experiments across key benchmarks demonstrate that HyPeR achieves absolute improvements over the base model, with performance comparable to large-scale models, stressing the effectiveness of hybrid perception-grounded reasoning, particularly in noisy and multi-speaker scenarios.

1 Introduction

Recent Large Audio Language Models (LALMs) have made strides in audio understanding (Chu et al., 2024; Kong et al., 2024; Tang et al., 2024), with steady progress on challenging audio reasoning benchmarks (Sakshi et al., 2024; Ma et al., 2025b). Yet, their performance is dominantly capped by perceptual errors, where the models struggle with distinguishing environmental sounds, and accurately transcribing or interpreting speech. Although LALMs have further made notable progress in reasoning via Chain of Thought

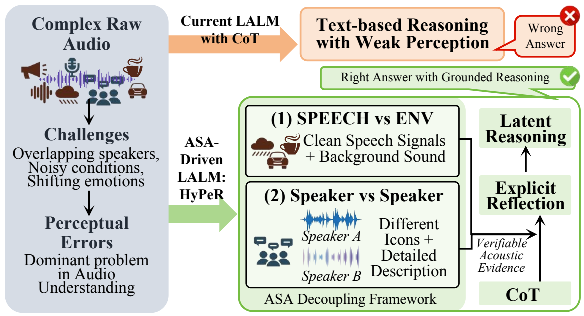


Figure 1: ASA-inspired Layered Decoupling for grounded audio comprehension. The model segregates background sound (ENV) from speech and distinguishes multiple speakers to generate verifiable acoustic evidence for LALMs.

(CoT) (Xie et al., 2025; Ma et al., 2025a) and reinforcement-learning (RL) post-training (Li et al., 2025a; Wu et al., 2025), the reasoning paths produced upon unreliable perceptions may hallucinate evidence and bring about bad comprehension in Audio Question-Answering (QA) (Yue et al., 2025). Moreover, current models often derive answers primarily from text-based reasoning without acoustic evidence, leading to weak audio grounding.

Previous research on audio grounding centered on Sound Event Detection with on- and off-set timestamps (Xu et al., 2021) and interval localization (Ghosh et al., 2024; Xiong et al., 2025), which brings about additional architectural complexity and extra inference time. Furthermore, it's hard for current LALMs to follow the routine since they may exhibit temporal misalignment (Kuan and Lee, 2025). To address these limitations, we focus on verifiable acoustic attributes and source-aware cues to improve audio grounding. Drawing inspiration from **Auditory Scene Analysis (ASA)**, the human brain processes complex soundscapes through layered decoupling pathways (Bregman, 1994; Michelsanti et al., 2021), effectively segregating the background sound (ENV) from the foreground one (SPEECH) and distinguishing multiple speakers before performing high-level semantic synthesis, as shown in Figure 1.

However, directly applying LALMs to background sound recognition remains unsatisfactory in practice. Specialized audio–text alignment models (e.g., CLAP (Elizalde et al., 2023, 2024; Ghosh et al., 2025; Niizumi et al., 2024)) report mean Average Precision (mAP) values below 50% on FSD50K, a multi-label audio tagging dataset, while Qwen2-Audio only achieves 15% mAP in our experiment. To address this gap, we introduce **PAQA**, a dataset specifically designed to benchmark and facilitate this decoupling. PAQA focuses on two core disambiguations: (1) **Speech vs. Environment**: isolating linguistic signals from non-speech interference; and (2) **Speaker vs. Speaker**: resolving multi-party attribution to recover conversational dynamics. PAQA contains 7,470 multiple-choice Audio-QA pairs, each enriched with structured annotations, including background-music separation, speaker analysis, and multi-turn reflections. By recording both internal acoustic cues and final responses, PAQA forces the model to ground its reasoning in explicit perceptual evidence.

To better detect and ground perceptual cues and acoustic attributes, we propose **HyPeR**, a two-stage **Hy**brid **Pe**rception-**R**easoning framework that unifies explicit reflective reasoning with implicit latent computation. Explicit Perception in Stage I involves Supervised Fine-Tuning (SFT) on PAQA to teach the model to imitate human-like layered auditory decomposition. Nevertheless, we observe that the generated CoT often remains imprecise when describing certain acoustic attributes (e.g., tone, pitch, background noise texture, and paralinguistic emotion). Inspired by Goyal et al. (2024), we mimic the “think before speak" pattern, and introduce a <PAUSE> special token that enables the model to perform latent reasoning based on Group Relative Policy Optimization (GRPO) before committing to verbal descriptions of difficult acoustic attributes. Moreover, we empirically find that when the model is about to generate tokens related to the acoustic keyword set, the token selection confidence is often lower. To better place the <PAUSE> token, we propose a sliding-window group confidence (Fu et al., 2025) to detect locally unreliable spans during generation. The reward function is designed for audio grounding and jointly balances answer correctness, reasoning consistency, and format compliance. Our experimental results on PAQA and other benchmarks demonstrate that HyPeR significantly reduces perceptual errors and achieves strong performance on complex audio understanding and reasoning tasks, particularly in noisy speech and multi-speaker scenarios.

Our contributions are summarized as follows:

- We focus on the Perception-Grounded Audio Understanding and redefine the reasoning of LALMs from a direct audio-to-text mapping to CoT with explicit acoustic grounding on environment sound and multi speakers based on Auditory Scene Analysis.
- We introduce PAQA, a novel benchmark designed to operationalize this hierarchical reasoning, with stepwise reasoning and reflection annotations across multi-speaker QA, noisy speech translation, and environment-centric QA, intended to suppress shortcut learning and promote acoustic grounding.
- We propose HyPeR, a hybrid framework that unifies explicit reflection with latent reasoning, with pause token detecting acoustic attributes. By employing a GRPO-based reinforcement learning strategy with multi-dimensional rewards (accuracy, consistency, and grounding), HyPeR effectively bridges the perception-reasoning gap.

2 Related Works

2.1 Large Audio–Language Models (LALMs)

Early LALMs such as Qwen2-Audio(Chu et al., 2024), Audio Flamingo(Kong et al., 2024), and SALMONN(Tang et al., 2024) advanced ASR, but remained fragile in real-world reasoning tasks involving multi speakers and non-stationary noise. More recent omni-/speech-native systems broaden the interface beyond transcripts with end-to-end audio generation such as OpenAI’s GPT-4o Audio models(OpenAI), and Gemini 2.5 Pro(Kavukcuoglu, 2025). However, on-demand CoT in Audio Flamingo 3(Goel et al., 2025a) and structured CoT in Audio-Reasoner(Xie et al., 2025), yet models often reverted to transcript shortcuts whenever acoustic evidence was difficult to verbalize. Recent work (Ghosh et al., 2024; Xiong et al., 2025) has therefore shifted toward architectural audio evidence alignment and multi-representation fusion, but brings about additional architectural complexity and extra inference time. To address these limitations, we release a structured dataset that couples multi-speaker and background-rich audio, explicitly guiding LALMs to ground decisions in acoustic rather than pure text.

2.2 Explicit Reasoning in LLMs

In LLMs, structured reasoning through CoT, reflection, and RL post-training has yielded consistent gains beyond supervised fine-tuning (SFT) (Guo et al., 2025; Team et al., 2025). While Vision-R1 (Huang et al., 2025) and Video-R1 (Feng et al., 2025) extended RL-based reasoning to overthinking suppression. In audio, GRPO-style RL underlies R1-AQA and Omni-R1 (Shao et al., 2024; Li et al., 2025a; Zhong et al., 2025), with mixed evidence on whether RL alone suffices. More recent approaches (Wen et al., 2025; Wu et al., 2025; Li et al., 2025b; Jin et al., 2025) highlight that objectives should reward useful and concise reasoning rather than verbosity. In this work, we instead unify explicit, audio-grounded reasoning with reflection, operationalized through a multi-term reward that enforces correctness and conciseness.

2.3 Implicit Latent Reasoning and Pause-Triggered Computation

Complementary to explicit rationales, implicit computation allocates additional internal processing before token emission. Learned <pause> tokens can trigger silent forward passes (Goyal et al., 2024), echoing earlier adaptive-computation approaches(Graves, 2017; Banino et al., 2021) that learn instance-dependent halting policies. To our knowledge, such latent computation has not been systematically validated in audio–language reasoning. Our contribution is to extend <pause> to LALMs and couple it with a lowest-group-confidence (LGC) controller: when confidence drops on acoustically inexpressible cues, HyPeR diverts into a short, budgeted latent stream and can abort tail trajectories under severe uncertainty.

3 Data with Audio Layered Decoupling

3.1 ASA-Inspired Taxonomy

To bridge the gap between raw acoustic signals and high-level reasoning, we introduce PAQA, which is designed to supervise the decoupling process itself, providing explicit "Perceptual Traces" based on Auditory Scene Analysis (Bregman, 1994). We further analyze Qwen2-Audio's bad cases on the CoTA (Xie et al., 2025) benchmark and identify two major challenges.

Level 1: Speech vs. Environment (S-E) To prevent the model from misattributing background interference as conversational evidence, we synthesize complex auditory scenes using MUSAN (Snyder et al., 2015) and FSD50K(Fonseca et al., 2021). For a speech clip s and an environmental noise n, we apply RMS-normalization and mix them with a dynamic SNR range of [0,20] dB. Crucially, each item is annotated with an Environment Tag (e.g., *"Background: Rain and distant traffic"*), forcing the model to distinguishing speech and non-speech during the reasoning phase.

Level 2: Speaker vs. Speaker (S-S) To resolve multi-party conversational structures, we annotate speaker turns using a structured format. To ensure the model performs true Speaker Attribution rather than shortcutting via global transcripts, we introduce the Quote-Presence Test (QPT). QPT measures the alignment between the model's attributed speaker segments and the raw ASR output(checked by Qwen3-ASR). We filter out items with QPT<0.85 to ensure the reasoning is strictly grounded in the temporal sequence of the audio. The alignment is formulated as:

$$\mathrm{QPT} = \frac{1}{M}\sum_{i=1}^{M} \max_{1\leq j\leq N} \phi(\hat{s}_i, \hat{a}_j), \qquad (1)$$

where $\hat{s}$ and $\hat{a}$ denote the normalized strings of attributed sentences and ASR snippets, respectively. $\Phi(\bullet)$ computes the fuzzy overlap ratio (SeqRatio) between two strings.

3.2 Data Collection & Statistics

In natural conversation, speakers frequently self-monitor and revise their utterances. Building on prior work showing that reflection-driven self-correction improves model performance in reasoning tasks (Shinn et al., 2023; Madaan et al., 2023; Wang et al., 2023), we adopt a reflection-augmented pipeline for complex audio understanding. Concretely, a lightweight baseline model first generates an initial <RESPONSE> for each audio QA item, as illustrated in Figure 2. We then automatically detect errors, such as option mismatches, speaker attribution mistakes, hallucinated content inconsistent with ASR transcripts, or misinterpretation of noise cues. Finally, we prompt the model to produce a grounded diagnostic analysis <REFLECT> with manual check. This analysis explicitly references <BGM>, <SPEAKER>, and <ASR> to explain the failure and localize the supporting evidence. Conditioned on this analysis, the model is guided to generate a corrected <FINAL_ANSWER>. For training, we store the triplet (<RESPONSE>, <REFLECT>, <FINAL_ANSWER>), which provides explicit reflection supervision and, from each original audio

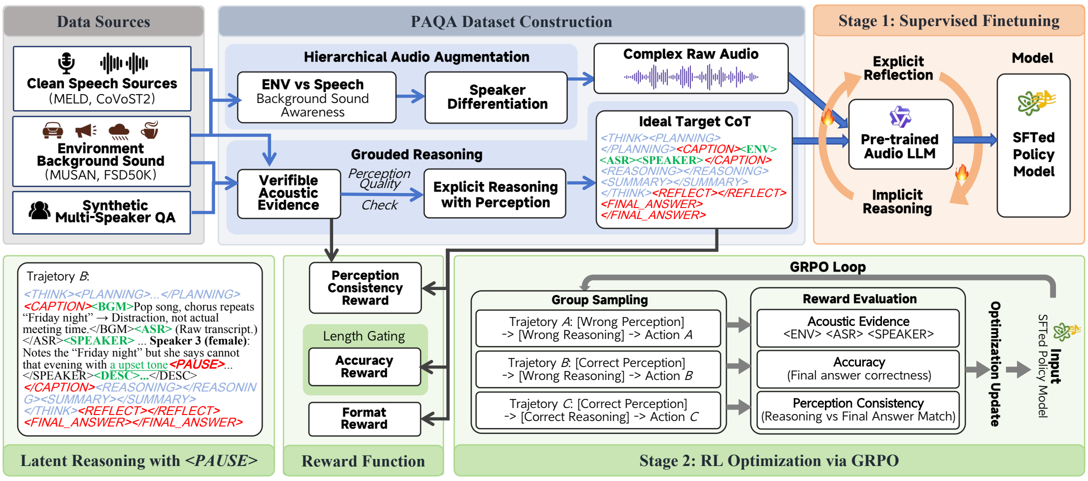


Figure 2: An overview of our framework HyPeR. First, we collected the PAQA dataset, with complex audio, and annotated perceptual information such as background sound and ASR transcript. Secondly, we fine-tuned on PAQA as the policy model in GRPO. The reinforcement learning mechanism includes latent reasoning with pause tokens, along with composite reward rules to improve performance.

item, yields an additional corrected example, effectively doubling the supervised data while enriching them with interpretable, perception, grounded self-correction signals. For detailed analysis and prompt template, see Appendix E.

4 Method

4.1 Overall Architecture

To bridge the gap between low-level acoustic perception and high-level audio-linguistic reasoning, we propose **HyPeR**, a unified Hybrid Perception-Reasoning framework that mimics the human brain's hierarchical processing of auditory scenes. Given an audio input X_a and a textual query Q, HyPeR aims to generate a logically grounded response Y. We decompose this into a two-stage hierarchical process: Explicit Perceptual Reflection and RL-driven Latent Reasoning.

We first enhance the model's perception through SFT on our PAQA dataset. The model is trained based on Qwen2-Audio to generate a structured reasoning chain that explicitly performs layered decoupling: first identifying the acoustic environment (Speech vs. Environment) and then resolving speaker dynamics (Speaker vs. Speaker). These traces, encapsulated within <REFLECT> tags, serve as the "logical grounding" for the final answer. Besides, recognizing that non-textualizable acoustic nuances (e.g., subtle prosodic shifts or overlapping textures) are difficult to describe explicitly, we introduce the <PAUSE> token. During the RL stage, the model learns to autonomously abort the trajectory when it encounters lower confidence. This allows dynamic latent reasoning, where the model allocates additional internal computation to refine its latent states before generating perceptual traces or the final response.

4.2 Stage I: Explicit Perception (SFT)

In this stage, the model is trained via Supervised Fine-Tuning (SFT) on the PAQA dataset to imitate human-like auditory decomposition. Following a structured reasoning pipeline, the model generates an explicit trace T consisting of four sequential components: (1) Planning (P): Outlining the logic required to address the query. (2) Captioning (C): Extracting multi-modal information, especially multi-layered acoustic features, including environment (<ENV>), speaker dynamics (<SPEAKER>), and speech content (<ASR>). (3) Reasoning (R): Performing step-by-step analytical deduction based on P and C. (4) Summary (S): Synthesizing the reasoning into a concise internal conclusion. (5) Reflection (R'): Producing a transparent analysis of background sound and speaker, and reflection that allows for direct inspection of the summary to a better answer. This process is formalized in Eq.2.

$$\begin{aligned}
P &\sim f_\theta(\mathbf{X}_a, \mathbf{Q}), \\
C &\sim f_\theta(\mathbf{X}_a, \mathbf{Q}, P), \\
R &\sim f_\theta(\mathbf{X}_a, \mathbf{Q}, P, C), \\
S &\sim f_\theta(\mathbf{X}_a, \mathbf{Q}, P, C, R), \\
R' &\sim f_\theta(\mathbf{X}_a, \mathbf{Q}, P, C, S).
\end{aligned} \tag{2}$$

The explicit trace $T = \{P, C, R, S, R'\}$ serves as the logical perceptual grounding for the final answer. We aim to teach the model to generate its responses in a specific, structured format, it lays the groundwork for the subsequent reinforcement learning phase. The optimization goal of this stage is the standard cross entropy loss in Eq.3.

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{|\mathbf{T}|} \log P(t_i \mid \mathbf{X}_a, \mathbf{Q}, \mathbf{T}_{<i}) \quad (3)$$

4.3 Confidence-based Transition Gating

After generating the explicit trace T, HyPeR evaluates whether the acoustic information has been sufficiently resolved. Audio streams contain a host of non-verbal cues, such as speaker intonation, overlapping speech, and ambient noise, that are often difficult to fully articulate in explicit text. We found a connection between the reasoning trace's lower confidence score and non-verbal cues. Therefore, we consider the Lowest Group Confidence (LGC) metric C_t at each decoding step t. Each token t is linked to a sliding window group K_i, consisting of n previous tokens. In particular, we identify its bottom 15% group confidence. For each window, we compute a normalized mean probability:

$$C_{K_i} = \frac{1}{|K_i|} \sum_{t \in K_i} C_t, \quad (4)$$

where $|K_i|$ is the number of tokens in group K_i. The LGC of the trajectory is then defined as the minimum of these window confidence scores, $\text{LGC}(\mathbf{y}) = \min_{k=1,\ldots,K} C_{K_i}$. This definition emphasizes the weakest local segment within the reasoning trajectory: even a small cluster of highly uncertain tokens can significantly reduce LGC, making it a sensitive indicator of detecting local reasoning collapse, a phenomenon effectively demonstrated by Fu et al. (2025).

When the LGC falls into the intermediate ambiguity range $(\tau_{abort}, \tau_{pause}]$, the model triggers a "Think-Before-Speak" reasoning step. If LGC drops below τ_{abort}, the model autonomously aborts the trajectory to prevent unproductive reasoning loops or hallucinations, significantly accelerating inference by pruning unpromising paths.

4.4 Latent Reasoning with Pause Token

During the initial phase of Stage II training, we introduce a keyword-based heuristic to calibrate the model's sensitivity to acoustic nuances. We maintain a keyword set K="tone", "pitch", "noise", "emotion", … representing non-textualizable cues. Whenever a word $w \in T$ appears in the recent context, we apply a positive logit bias $\beta_{ac} > 0$ to the <PAUSE> token, as shown in Figure 6:

$$\ell_{\texttt{<PAUSE>}} \leftarrow \ell_{\texttt{<PAUSE>}} + \beta_{\text{ac}} \cdot \mathbb{I}\big[\exists w \in \mathcal{K}\big] \quad (5)$$

This mechanism serves as a cold-start prior for the threshold τ_{abort}, encouraging the model to allocate latent computation specifically when the explicit text involves speech-only cues.

When a pause is triggered at step t, the model emits a <PAUSE> special token and generates a sequence of latent tokens $\hat{\mathbf{z}}_{1:L}$. Crucially, these tokens function as a non-volatile computational cache; they are not surfaced in the final visible output and are explicitly excluded from the gradient calculations during the generation of the final response to maintain efficiency. Their function is only to iteratively update and refine the model's internal hidden state H_t, enabling a deeper, more grounded processing of complex audio features before resuming the generation of visible tokens. The relationship between the full internal sequence $\tilde{\mathbf{y}}$ and the visible output y_{vis} is formalized as:

$$\tilde{\mathbf{y}} = \mathbf{y}_{1:t^\star} \oplus \texttt{<PAUSE>} \oplus \hat{\mathbf{z}}_{1:L}, \; \mathbf{y}_{\text{vis}} = \mathbf{y}_{1:t^\star} \quad (6)$$

The architecture ensures the model "thinks" internally as it processes intricate auditory scenes, effectively bridging the gap between low-level acoustic perception and high-level text reasoning.

4.5 Stage II: GRPO-based RL Post-Training

While Supervised Fine-Tuning (SFT) in Stage I establishes a structural foundation for auditory decomposition, its efficacy is inherently limited by the nature of imitation learning. To optimize the model's internal reasoning ability, we introduce a second stage of optimization using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) from the SFT checkpoint as the reference policy π_{ref} frozen. We generate groupwise rollouts, compute $R(\mathbf{z})$ via (9), and update π_θ with GRPO (Shao et al., 2024). We partition rollouts by task group $g \in \{\text{PAQA}, \text{AVQA}\}$. For each trajectory i within a group, we compute the relative advantage to reduce variance:

$$\tilde{R}^{(i)} = R^{(i)} - \frac{1}{m_g} \sum_{j \in g} R^{(j)}, \quad (7)$$

where m_g is the number of samples in the group.

To specifically address the "thinking" process regarding non-textual audio cues, we utilize the keyword set K (e.g., "tone", "pitch", "noise") as a cold-start prior. In early RL iterations, these keywords provide initial guidance on acoustic sensitivity by influencing the gating threshold τ_{pause}. Crucially, we incorporate the Lowest Group Confidence (LGC) metric C_t into the advantage calculation. The LGC serves as a proxy for the "logical weakest link" in a reasoning trajectory. For a trajectory i with a raw task reward r_i^{task} (encompassing accuracy, formatting, and consistency), the weighted advantage A_i is defined as:

$$A_i = w_i \cdot (r_i^{\text{task}} - \bar{r}), \tag{8}$$

where $w_i = clip(std(LGC(y)))$ is a standardized weight derived from the trajectory's LGC. Here, $w_i = 0$ for trajectories that fall below the τ_{abort} threshold, effectively pruning unpromising or unstable reasoning paths during optimization.

4.6 Multi-Objective Reward Function

To ensure the model not only produces accurate answers but also generates interpretable, perception-grounded reasoning, we design a composite reward function R. It is defined as a weighted sum of four specialized components:

$$\begin{aligned} R =& w_{\text{acc}}\, \mathcal{R}_{\text{acc}} \;+\; w_{\text{cons}}\, \mathcal{R}_{\text{cons}}(\hat{y}, \hat{y}_{\text{CoT}}) \;+ \\ & w_{\text{fmt}}\, \mathcal{R}_{\text{fmt}} \;+\; w_{\text{len}}\, (\mathcal{R}_{\text{acc}} \times \mathcal{R}_{\text{len}}), \end{aligned} \tag{9}$$

where $\mathcal{R}_{\text{acc}}$ and $\mathcal{R}_{\text{fmt}}$ provide the fundamental supervision for task completion, while $\mathcal{R}_{\text{cons}}(\hat{y}, \hat{y}_{\text{CoT}})$ and $\mathcal{R}_{\text{len}}$ serve as perceptual and structural regularizers to stabilize the learning of the hybrid reasoning process.

4.6.1 Accuracy and Format Rewards

The Accuracy Reward ($\mathcal{R}_{\text{acc}}$) is a binary signal $\mathbf{1}[\hat{y} = y]$. We prioritize extracting $\hat{y}$ from the <FINAL_ANSWER> tag, with a fallback to the <RESPONSE> tag to ensure robustness during early RL stages. The Format Reward ($\mathcal{R}_{\text{fmt}}$) addresses the reward sparsity inherent in complex structural requirements. To prevent "gradient collapse" where the model fails to receive any signal due to strict schema violations, we adopt a progressive format shaping strategy. We reward a "weak format" (correct <THINK> and <RESPONSE> sequence) with a base score, while the "strict format" (inclusion of specific environment and speaker tags) is implicitly incentivized through the consistency rewards described below.

4.6.2 Perceptual Consistency Reward

To enforce the "perception-grounded" nature of our framework, $\mathcal{R}_{\text{con}}$ regularizes the reasoning chain along three acoustic-logical axes:

BGS Robustness. To eliminate illusions where the model treats background sound as causal evidence for speech-related questions, we define a background sound gate ∇_{bgs}. If the reasoning chain invokes environmental cues (e.g., "the background music suggests...") as a causal basis for linguistic content, ∇_{bgs} is set to 0; otherwise, it is 1.

Speaker-ASR Fidelity. Within the <THINK> block, we extract speaker-attributed quotes $S = s_i$ and verify them against the raw ASR transcript $A = a_j$. We define the fidelity score ∇_{fid} as:

$$r_{\text{fid}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \max_{a \in \mathcal{A}} \phi(\hat{s}, \hat{a}), \tag{10}$$

where ϕ is the character-level Levenshtein similarity. This ensures that the model's "perception" is strictly anchored to the acoustic evidence rather than hallucinated text.

Reasoning-Answer Alignment. We reward the agreement between the model's internal conclusion $\tilde{y}$ in CoT and its final answer $\hat{y}$.

The final consistency reward is:

$$\mathcal{R}_{\text{cons}} = \nabla_{\text{bgs}} \cdot \left(\lambda_{\text{fid}} r_{\text{fid}} + \lambda_{\text{align}} r_{\text{align}}\right). \tag{11}$$

4.6.3 Length Shaping via Correctness Gating

To prevent "reasoning collapse" (too short) or "superficial verbosity" (too long), we introduce $\mathcal{R}_{\text{len}}$, which is only activated when $\mathcal{R}_{\text{acc}} = 1$. We use a piecewise-linear function with a penalty for completions exceeding T_{max} tokens or failing to reach T_{min} tokens. Crucially, any content generated after the </FINAL_ANSWER> tag results in a zeroed length reward to encourage clean termination.

5 Experiments

5.1 Implementation Details

All experiments fine-tune the same pretrained backbones (Qwen2-Audio-7B-Instruct), using the framework introduced by Li et al. (2025a). Training is conducted with a batch size of 1 per GPU, with by 2 gradient accumulation steps, resulting in an effective total batch size of 16. We adopt a learning rate of $1e-6$, a temperature of 1.0, and configure the GRPO to sample 8 responses per group with a KL coefficient β of 0.1. For models incorporating pause a latent thinking mechanism, we set $\tau_{pause} = 0.5$ and allow up to 3 pauses per sequence with 64 thinking tokens each, plus $\tau_{abort} = 0.05$ for think token containment.

5.2 Benchmarks and Metrics

We evaluate six configurations: **SFT**, standard fine-tuning; **GRPO-Nothink**, GRPO post-training without <REFLECT> or <PAUSE>; **GRPO+CoT**, GRPO enhanced with thinking before the answer (in the weak format of <THINK><ANSWER>); **GRPO+ExpCoT**, GRPO enhanced with explicit <THINK> (including <REFLECT>) but no <PAUSE>; **Ours (HyPeR)**, GRPO enhanced with the explicit schema and <PAUSE>; and **External Baselines** including GPT-4o Audio(Jaech et al., 2024), Gemini 2.5 Flash(Comanici et al., 2025), Audio-Flamingo-3(Goel et al., 2025b), OmniVinci(Hanrong Ye, 2025), Qwen2.5-Omni(Xu et al., 2025), and existing LALM reasoning frameworks like Audio-Reasoner (Xie et al., 2025), Audio-CoT (Ma et al., 2025a) and Audio-Thinker (Wu et al., 2025) (all trained on Qwen2-Audio-7B).

We use PAQA (train set) for supervised finetuning. For RL training, we utilize 30,000 augmented samples generated upon the AQVA (Yang et al., 2022) dataset, with each response reformulated into a `<think>...</think><answer>...</answer>` reasoning–answer structure. Models are evaluated on several benchmarks, **PAQA Test** (“MSQA-hard" for the subset of QA with >3 speakers, “ENVQA-hard" for the subset with background sound under SNR=5dB), **MMAU** (Sakshi et al., 2024), and **MMAR** (Ma et al., 2025b). The results are listed below and in the Appendix.C.

5.3 Direct LALM Percepting Underperforms

To evaluate LALM’s perception ability, we first use models directly recognizing background sound on FSD50K, a multi-label sound event classification benchmark, and calculate Word Error Rate (WER) and Character Error Rate (CER) based on the transcripts generated in the explicit reasoning on the PAQA test set. Qwen2-Audio achieves only 14.7% mAP on FSD50K, far below the audio–text alignment model CLAP23(Elizalde et al., 2023) ’s 50%, and poor for direct generation in multi-label environmental sound tagging. HyPeR narrows the gap to 43.6% and achieves a remarkably low WER of 1.65% and CER of 1.61%, demonstrating that our model’s reasoning is grounded in more accurate perception, ruling out hallucination.

5.4 Main Results

We evaluate HyPeR against multiple LALMs on MMAU Test-mini and MMAR. As shown in Table 2, our method achieves performance competitive with large-scale models on complex audio understanding tasks, particularly in speech.

Table 1: Results on FSD50k sound event classification and WER, CER in the explicit reasoning on the PAQA.

Model	FSD50k	WER	CER
HyPeR (Ours)	0.436	**0.781**	**0.623**
Qwen2-Audio (base)	0.147	0.869	0.779
CLAP23	**0.486**	23.071	24.801

Table 2: Performance on MMAU Test-mini (Sakshi et al., 2024) and MMAR (Ma et al., 2025b).

Method	MMAU Test-mini↑				MMAR↑
	Sound	Music	Speech	Avg.	Avg.
Gemini 2.5	67.97	62.28	62.76	64.30	**66.80**
GPT-4o	61.56	56.29	66.37	61.40	63.50
Audio-Flamingo-3	**79.58**	73.95	66.37	**73.30**	58.50
OmniVinci	73.65	**78.68**	66.97	73.10	58.30
Qwen2.5-Omni	78.10	65.90	**70.60**	71.50	56.70
Qwen2-Audio	61.26	53.59	48.05	54.30	30.00
+SFT	62.76	44.61	55.86	54.41	40.90
+GRPO	68.17	61.38	60.66	63.40	45.40
+GRPO +ExpCoT	75.07	58.98	63.66	65.90	48.20
Ours (HyPeR)	75.67	62.27	64.26	67.40	55.50
Audio-CoT	62.16	55.99	56.16	58.10	31.67
Audio-Reasoner	60.06	64.30	60.70	61.71	36.71
Audio-Thinker	76.88	62.87	64.26	68.00	52.00

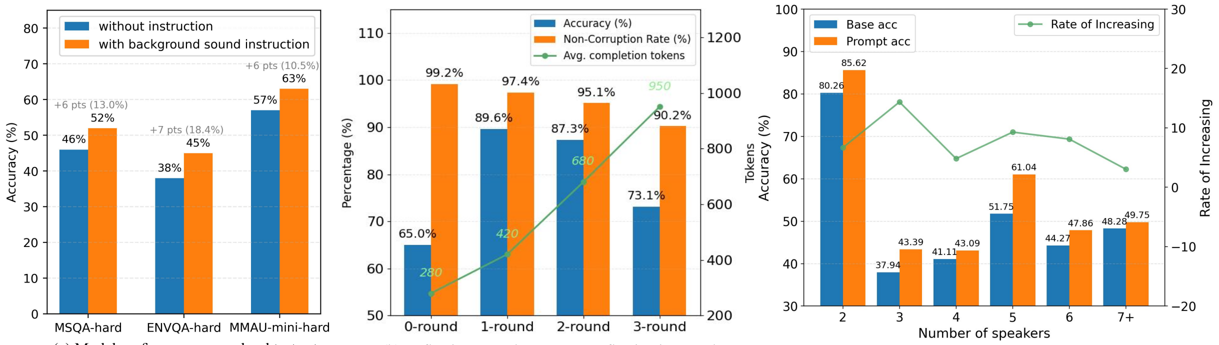
RL vs. SFT While GRPO without reasoning (No-Think) improves accuracy, the most substantial gains occur when combining Explicit Perceptual Traces (Stage I) with Implicit Latent Computation (Stage II). HyPeR offsets the domain shift observed in the Music subset during SFT, suggesting that RL helps the model adapt its perceptual boundaries to diverse acoustic scenes.

Pause mechanism works. The implicit reasoning enabled by <PAUSE> tokens during ambiguous acoustic phases is particularly effective in complex audio environments, especially on naturally occurring mixed-modality audio(MMAR +25.5). Notably, it improves the Music subset, offsetting the bad performance of just finetuning. More detailed analyses are provided in Appendix C.3.

5.5 Ablation Study

5.5.1 Robustness to ENV and Multi-Speaker

Background Sound As shown in Fig.3(a), we evaluated that once the model is informed of background sound (one parameter of the prompt), it can correctly detect if that “noise” is unrelated to the main dialogue content. The introduction of background sound in the original audio leads to measurable degradation of zero-shot performance.



(a) Model perform poorer on hard test set with background sound disturbing. But sound instruction recovers accuracy.

(b) Reflection Rounds. One-turn reflection is enough. More rounds of reflection tend to have longer reasoning answer but more frequent corruption.

(c) Ablation study on multi-speaker problems. Orange column means adding speaker information in the prompt. Information of speaker analysis leads to better performance.

Figure 3: Comparison between different audio situations.

However, this drop is substantially mitigated while explicit "ignore background sound" prompts are provided. This validates that our reflection step substantially improves accuracy. In Fig. 3(b), we further compare the effect of varying numbers of reflection turns, moving from 0 to 1 round, which yields a large accuracy enhancement. However, adding more rounds leads to "overthinking" and worse results, suggesting that longer text-based reasoning is useless.

Multi Speakers Overall, recognizing the environment sound improves accuracy, which is consistently beneficial across all speaker counts. The base model is strong with 2 speakers (80.26%), but drops sharply with more speakers. This pattern matches the intuition that more speakers introduce attribution and coreference errors. For 7+ speakers, the improvement is modest, suggesting that richer cues (e.g., explicit diarization tags, role summaries, or brief scene summaries) are likely needed.

5.5.2 Reward Function

As shown in Table 3, we compare HyPeR and GRPO without Consistency Reward and length shaping respectively. The results demonstrate that the consistency reward ensures the model's logic is strictly grounded in the ASR and environment sound, leading to a 4.2% gain in overall reliability.

Table 3: Ablation of rewards of Accuracy (Acc.) and Consistency (Cons.) on PAQA test set.

Config	**Acc.**	**Con.**
Full Reward (HyPeR)	**68.4**	**91.2**
w/o Consistency Reward ($\mathcal{R}_{\text{con}}$)	64.2	78.5
w/o Length Shaping ($\mathcal{R}_{\text{len}}$)	67.1	89.4

5.5.3 Do PAUSE Tokens Enable Latent Reasoning in Audio?

To investigate whether the <PAUSE> tokens facilitate genuine latent computation or merely prolong decoding, we analyze the evolution of the model's top-layer hidden states h_t during the pause phase by tracking two metrics across pause indices i: (1) **Cosine Similarity to Answer** $\cos(h_{pause,i}, h_{ans})$, measuring how much the representation aligns with the final correct output; and (2) **Step-wise Displacement** $\|\Delta h\| = \|h_i - h_{i-1}\|$, quantifying the magnitude of state updates. As shown in Table 4, the displacement $\|\Delta h\|$ remains significantly above zero, confirming that the hidden states are undergoing active transformation rather than staying stagnant. While initial pauses may involve exploratory shifts, the trajectory eventually converges towards the answer embedding, suggesting that the model uses the latent space to refine its internal evidence before generating the final token.

Table 4: Evolution of hidden states across sequential PAUSE tokens (Averaged over 100 samples).

Metric/PAUSE Token	#1	#2	#3	Final Ans
Avg. Cos-Sim to Ans	0.47	0.51	0.62	0.73
State Displacement $\|\Delta h\|$	-	336.2	324.8	338.5
Trigger Freq. (per sample)	1.00	0.78	0.45	-

6 Conclusion

In this paper, we argue that improving audio understanding requires the base model to have audio grounding. Based on Auditory Scene Analysis, we focus on verifiable acoustic evidence and first introduce PAQA, a dataset that implements a layered decoupling strategy to separate speech from environmental interference and resolve multi-speaker attribution. Building upon this, we proposed HyPeR, a hybrid framework that unifies explicit perceptual reflections with implicit latent reasoning with GRPO-based <PAUSE> tokens. Experiments demonstrate that HyPeR significantly reduces perceptual errors and improves reasoning ability with evidence-constrained acoustic grounding.

Limitations

Despite the significant improvements achieved by HyPeR and PAQA, several limitations remain to be addressed in future work:

First, the introduction of the <PAUSE> token mechanism inevitably increases both training and inference latency. Although our proposed Abort Mechanism partially mitigates this, finding an optimal balance between reasoning depth and real-time responsiveness remains a significant challenge. Future work will explore more efficient latent reasoning architectures to minimize latency without sacrificing the robustness of audio grounding.

Besides, while our framework significantly improves audio understanding, it does not achieve SOTA results. However, HyPeR achieves highly competitive performance using only 7.4k high-quality, perception-grounded samples from the PAQA dataset, underscoring the superior data efficiency of our approach. Detailed analysis reveals that HyPeR's improvements are primarily driven by the logical alignment of speech and environmental sounds rather than simple category memorization.

Ethical Considerations

Regarding Data Privacy, all audio samples in the PAQA dataset are derived from publicly available sources with permissive licenses, and any potentially sensitive speech content has been manually screened and anonymized to protect individual privacy. The license of MUSAN is CC_BY 4.0, which permits free use for academic research and modification, and we have cited the work.

References

Andrea Banino, Samuel Ritter, and 1 others. 2021. Pondernet: Learning to ponder. In *ICML*.

Albert S Bregman. 1994. *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report.

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*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2024. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340. IEEE.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *Preprint*, arXiv:2503.21776.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence. *Preprint*, arXiv:2508.15260.

Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2025. Reclap: Improving zero shot audio classification by describing sounds. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025a. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

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. 2025b. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. *Preprint*, arXiv:2310.02226.

Alex Graves. 2017. Adaptive computation time for recurrent neural networks. *Preprint*, arXiv:1603.08983.

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.

Arushi Goel Wei Huang Ligeng Zhu Yuanhang Su Sean Lin An-Chieh Cheng Zhen Wan Jinchuan Tian Yuming Lou Dong Yang Zhijian Liu Yukang Chen Ambrish Dantrey Ehsan Jahangiri Sreyan Ghosh Daguang Xu Ehsan Hosseini-Asl Danial Mohseni Taheri Vidya Murali Sifei Liu Jason Lu Oluwatobi Olabiyi Frank Wang Rafael Valle Bryan Catanzaro Andrew Tao Song Han Jan Kautz Hongxu Yin Pavlo Molchanov Hanrong Ye, Chao-Han Huck Yang. 2025. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *Preprint*, arXiv:2503.06749.

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*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Preprint*, arXiv:2503.09516.

Koray Kavukcuoglu. 2025. Gemini 2.5: Our most intelligent ai model. `https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/`. Accessed 2025-12-22.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *Preprint*, arXiv:2402.01831.

Chun-Yi Kuan and Hung-yi Lee. 2025. Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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

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

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. *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, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025b. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Preprint*, arXiv:2505.13032.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, Masahiro Yasuda, Shunsuke Tsubaki, and Keisuke Imoto. 2024. M2d-clap: Masked modeling duo meets clap for learning general-purpose audio-language representation. *arXiv preprint arXiv:2406.02032*.

OpenAI. Gpt-4o audio model (gpt-4o-audio-preview) | openai api documentation. `https://platform.openai.com/docs/models/gpt-4o-audio-preview`. Accessed 2025-12-22.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

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. *Preprint*, arXiv:2410.19168.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan

Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *Preprint*, arXiv:1510.08484.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *Preprint*, arXiv:2007.10310.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. 2025. Sari: Structured audio reasoning via curriculum-guided reinforcement learning. *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. *Preprint*, arXiv:2508.08039.

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

Zhen Xiong, Yujun Cai, Zhecheng Li, Junsong Yuan, and Yiwei Wang. 2025. Thinking with sound: Audio chain-of-thought enables multimodal reasoning in large audio-language models. *arXiv preprint arXiv:2509.21749*.

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*.

Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. 2021. Text-to-audio grounding: Building correspondence between captions and sound events. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 606–610. IEEE.

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.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. 2025. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *Preprint*, arXiv:2505.20256.

A Details of Data Collection

A.1 Synthetic Audio with Background Sound

Following this, we further analyzed erroneous predictions of Qwen2-Audio on the MMAU benchmark. As shown in Fig.8(b), we compared fine-tuning trajectories on the MSQA dataset with and without ASR-augmented data. The results reveal that models trained with ASR supervision exhibit substantially longer response lengths, which we interpret as a proxy for deeper and more structured reasoning ability. This finding suggests that integrating ASR data into training not only improves transcription accuracy but also enhances the reasoning capacity of audio-language models. Therefore, in the first stage of fine-tuning, we deliberately incorporated the ASR-enriched data described in the previous section to further consolidate the model's ASR capability as a foundation for downstream reasoning.

Moreover, we processed the audio with MUSAN(Snyder et al., 2015), which satisfies target 10 dB SNR, according to

$$\text{SNRdB} = 10 \log 10 \left(\frac{P_s}{P_{n,\text{scaled}}} \right) = 10.$$

Let $P_s = \frac{1}{T} \sum_t s_t^2$ and $P_n = \frac{1}{T} \sum_t n_t^2$. The background gain is

$$k = \sqrt{\frac{P_s}{P_n \cdot 10^{\text{SNR}_{\text{dB}}/10}}} = \sqrt{\frac{P_s}{P_n \cdot 10}}.$$

A.2 Audio Question-Answering with Multi Speakers

We use the subset of Multi-Speaker Dataset in CoTA (Xie et al., 2025), which is generated by TTS to navigate intricate speaker interactions. First, we generated diverse conversational texts with LLMs. Next, using timbres from LibriSpeech as prompts, we synthesized high-quality speech via the CosyVoice2 framework. Finally, these distinct speech samples were combined into a rich dataset.

B Data Statistics

A detailed case in shown in Figure 4.

The dataset supports a range of tasks, including multi-speaker QA, speech-to-text translation under noise, and environment-centric QA. An in-depth analysis of the final PAQA dataset is provided in Appendix A, while a detailed statistical overview is summarized in Table 5.

C Additional Results

C.1 Number of the Pause Tokens

Excessive pausing negatively affects performance(see Fig. 5), suggesting that it is suitable to set max pause token between 1 and 3.

C.2 Results on the test set of PAQA

We also evaluate on the test set of PAQA, on the category of multi-speaker and MELD (Xie et al., 2025), HyPeR performs the best. The results is listed in Table. 6.

Furthermore, under the challenging setting with background sound at SNR=5dB, a condition that considerably degrades most models, our HyPeR deteriorates the least, retaining state-of-the-art accuracy and consistency. This resilience is attributed to its pause-driven implicit reasoning and rewards aware of background sound/music.

C.3 Proper Response Length after Latent Reasoning

Though more stable during training, introducing pause-based latent tokens increases training time, raising max_pause_token from 1 to 3 roughly doubles training time. See more details in Fig.??. Therefore, we set a length reward in the design of whole reward function. We also observe some findings about the design of length-reward Sec. 4.6.3. Overall, the RL training progressed well, but there is often a clear performance drop about 200 steps. The instability can be attributed to the length-reward: during RL exploration, the model received higher scores for generating longer responses, but once a response exceeded 600 tokens, a linear decay penalty kicked in. In reaction, the policy abruptly shifted to producing shorter outputs; these truncated responses were often incomplete, leading to a format reward drop to zero and a reduction in accuracy reward to 0.5. Following this disruption, the training process gradually recovered and ultimately stabilized, indicating the policy capacity to adjust its generation in response to complex reward signals).

Overall, the RL training progressed well, but there is a clear collapse around 200 steps. The trigger was the length-reward design: during exploration, longer completions earned higher scores, but once a response exceeded 600 tokens, a linear decay penalty kicked in. The policy reacted by abruptly shortening completions to 200 tokens; these outputs were often incomplete, so the format

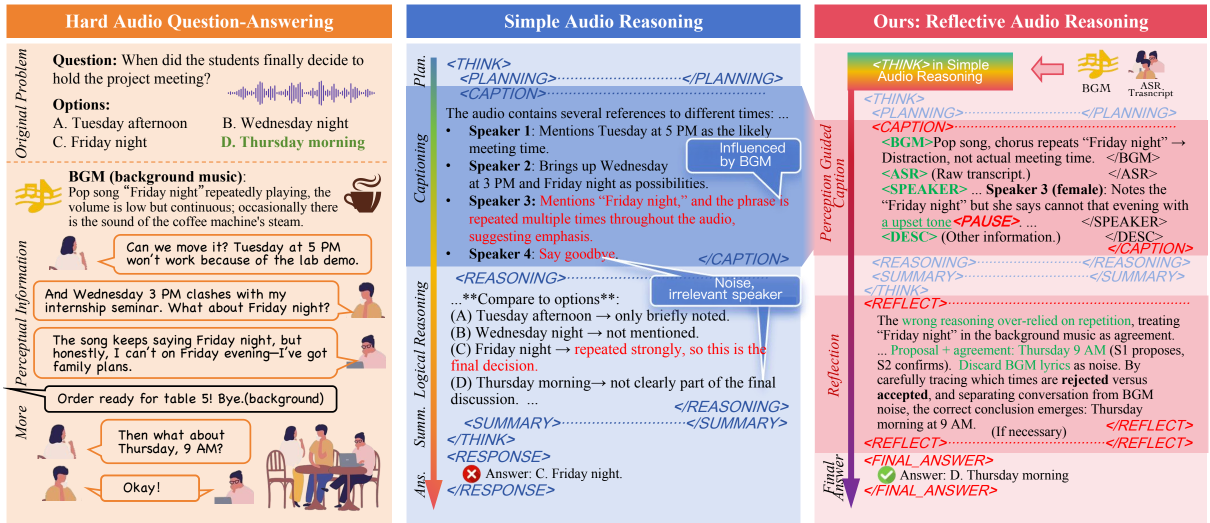


Figure 4: Case study.

Table 5: Dataset Source and Statistics. “MS" means whether there are multi speakers in the audio.

Dataset Source	Main Skills Learning	BGM Used	Quantity	Reflection	duration	MS
Multi-Speaker (Xie et al., 2025)	Multi-speaker Speech QA	Free Sound	2.9k	1.4k	264	✓
MELD (Poria et al., 2019)	Speech Emotion QA	Sound Bible	2.9k	1.4k	359	✓
CoVoST2 (Wang et al., 2020)	Speech-to-Text Translation	No	1.4k	No	72	×

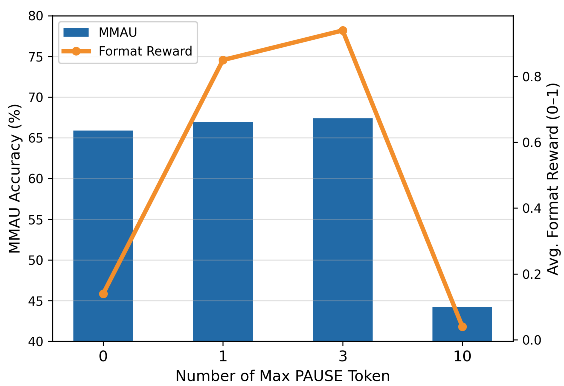


Figure 5: Abaltion study of #<PAUSE> tokens. Set max pause token as 1-3 is suitable.

Model	Multi-Speaker(hard) Acc.	Con. ↑	BGM-rich Acc. SNR=10	SNR=5
Qwen2-Audio	42.2	38.5	41.0	20.1
+SFT	46.2	41.5	44.0	31.2
+GRPO-NoThink	52.7	48.3	50.2	38.4
+GRPO-ExpCoT	61.5	58.7	60.8	47.6
Ours	**70.4**	**68.1**	**69.5**	**57.8**
Audio-CoT	50.6	46.9	48.3	35.0
Audio-Reasoner	56.8	52.7	55.9	41.8

Table 6: Evaluation on the test set of PAQA. Our model performs best in each category.

reward dropped to 0, and the accuracy reward fell to 0.5. After this shock, training recovered and stabilized, indicating the policy adapted to the length constraint.

D Case Study

As shown in Fig.4, the case highlights two failure modes: perceptual misbinding and salience-driven rationale drift. The naive system exhibits this by prioritizing the surface frequency of "Friday" while overlooking its negated polarity and the logical flow of the proposal-to-confirmation sequence. Conversely, the reflective controller rectifies this by enforcing evidence typing (differentiating background sounds from linguistic turns) and ensuring dialogue-act alignment, ultimately restoring causal fidelity to the acoustic evidence.

D.1 Explicit Reflection

Bad Case A

<THINK>
<PLANNING>
The user's question is: Based on the conversation, what are the speakers trying to assemble? I need to analyze the dialogue between the speakers to identify the item they are constructing. The audio provides the audio clip, and I should analyze it for details related to the assembly activity. </PLANNING>
<CAPTION>
The audio features a conversation among three individuals in what sounds like a workspace. They are discussing an instruction manual and the assembly process. - Speaker 1: Expresses frustration with the instruction manual, stating that it's difficult to distinguish pieces A and B.
- Speaker 2: Proposes that piece B has a slightly beveled edge, but concedes that it's similar.
- Speaker 3: Agrees that it looks like Bevel Edge and expresses hope that it's correct.
- Speaker 1: States it is Bevel Edge B, and declares they can complete the assembly.
- Speaker 2: Hope is right and doesn't want to misassemble again.
- Speaker 1: Confirms it is Bevel Edge B and asks everyone to take deep breaths and move on.

The core task is assembling an item where they are currently putting together parts, such as bevel edges. </CAPTION>
<REASONING>
1. Analyzing the Dialogue: The dialogue revolves around distinguishing parts (A and B) and confirming a beveled edge.
2. Identifying the Assembly Activity: The key element is Bevel Edge, suggesting the speakers are assembling something with beveled sides.
3. Evaluating Options: - (a) A table: While a table might require beveling, the central focus isn't just about assembling a table. - (b) A chair: Similarly, a chair might require beveling, but it isn't their primary focus. - (c) A bed: A bed typically features beveled edges, aligning perfectly with the conversation. This is the most likely scenario. - (d) A bookshelf: A bookshelf is less likely to have beveled edges.
4. Concluding the Item: Based on the bevel edges and the distinct focus on distinguishing them, it can be inferred that the speakers are assembling a bed. </REASONING>
<SUMMARY>
The speakers are differentiated by focusing on the Bevel Edge assembly of a bed. By eliminating other less relevant options through reasoning, the final conclusion is that they are assembling a bed. </SUMMARY>
</THINK>

Here is the reflection of this case.

Reflection of Bad Case A

<REFLECT> Evidence available from the dialogue is generic to many flat-pack builds: "piece A vs. B," "bevel(ed) edge," "don't misassemble," and "let's move on." No category-specific anchors are heard (e.g., for a bed: headboard/footboard/rails/slats; for a bookshelf: shelf/upright/back panel/bracket; for a table: table-top/apron; for a chair: backrest/seat/armrest). Given absent anchors, the safest policy is to avoid over-commitment. If forced to choose among the four, "bookshelf" is the most compatible with beveled front edges on shelves and frequent A/B part distinctions during shelf panel assembly. Confidence remains low; with an "uncertain" option, I would state insufficient evidence. </REFLECT>
<ANSWER>Bookshelf.</ANSWER>

D.2 Multi Speakers Tracking

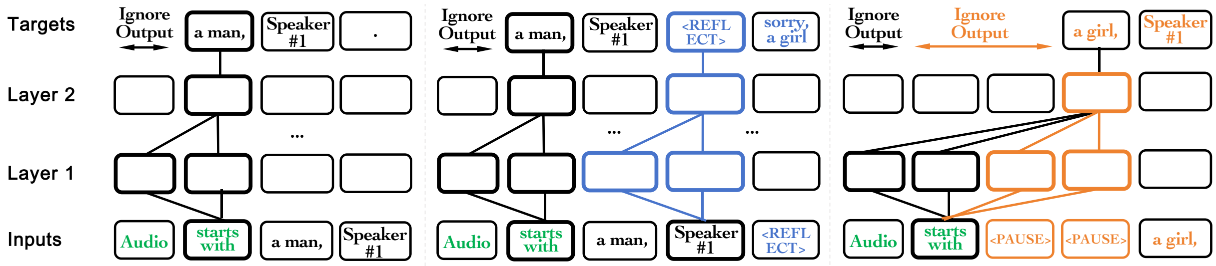
This case tests how well the model can follow a conversation in a realistic, noisy meeting. The audio includes a constant low hum from an air conditioner (HVAC) in the background, and the speakers often interrupt each other. These challenges make it difficult to follow the flow of the conversation. The simple model's failure wasn't about hearing the words; it was about failing to understand the process of the discussion. The reflective model succeeded because it built a step-by-step understanding of the conversation's logic. It correctly handled the disagreements and changes, proving it can reason reliably even in a noisy and dynamic group setting.

Bad Case B

[ASR excerpt] [S1] "Ship date is the 12th if QA passes." [S3] "QA won't finish by the 12th." [S4] "Set the launch to the 15th." [S2] "Not the 5th—I said the 15th." [S1] "Agreed."
[Question] What is the final launch date?
[Choices] A 5th · B 12th · C 13th · D 15th
[Audio context] Meeting room; steady HVAC hum as BGM; frequent interruptions.
[Ground truth] D
[Simple model response] B
[Reflection] <REFLECT> S3 explicitly negates the initial "12th" suggestion, S4 introduces a new proposal "15th," S2 clarifies that it is "15th" rather than "5th," and S1 provides team confirmation; taken together these turns establish the final launch date as the 15th.

D.3 Pause Mechanism

This case tests the model's ability to understand meaning that isn't in the words themselves, but in the way they are spoken (the "tone," or prosody). The audio is a phone call with background noise from rain and static (line hiss). The challenge is to see if the model can spot sarcasm or reluctance, even when the words seem positive. Therefore, this case highlights a major flaw in models



(a) Standard. (b) Reflective. (c) Pause-triggerred.

Figure 6: Standard vs. reflective-inference vs. pause-inference (and finetuning). Similar to Goyal et al. (2024), the rounded squares denote one Transformer operation (a self-attention and MLP) in a 2-layer Transformer. Any Ignore Output denotes that during inference, the corresponding output token is not extracted and thus, not fed back autoregressively.

that only read text. They can be easily fooled when a speaker's tone changes the meaning of their words (like in sarcasm or politeness). The reflective model's success shows it can listen, not just read. It integrates how something is said with what is said to find the true meaning, which is crucial for understanding human conversation.

Bad Case C

[ASR excerpt] [S1] "Can you present Friday?" [S2] "Yeah, Friday... great." (F0 downward; reduced energy; elongated final syllable) [S2] "Actually, I'm not available Friday evening." [S1] "Then Thursday morning?"
[Question] Is S2 willing to present on Friday?
[Choices] A Yes · B No · C Unsure · D Only afternoon
[Audio context] steady rain and mild line hiss
[Ground truth] B
[Simple model response] A
[Reflection] <REFLECT> ..."Yeah, Friday... great" <PAUSE> check tone/intonation. </PAUSE> The phrase "Yeah, Friday... great" is produced with a downward F0 contour and reduced energy consistent with reluctance or irony, and a subsequent turn states explicit unavailability on Friday evening; integrating prosodic and semantic cues indicates S2 is not willing to present on Friday.
</REFLECT>

In sum, the Multi-speaker Tracking matters speaker attribution ("who"), the Pause Mechanism addresses the transition from rapid, text-centric processing to a more computationally intensive, multi-modal analysis, and Reflection focuses on the ground truth by deploying targeted evidence re-querying, contextual anchoring to disambiguate local hypotheses (as in ASR N-best lists), and integrating conflicting cross-modal data. The convergence of these mechanisms allows the model to emulate human-like cognitive robustness in complex, ambiguous, and multi-participant acoustic environments, distinguishing its performance from that of a passive, deterministic system.

E Prompt Template

A Sample Case of Our PAQA Data

<THINK>
<PLANNING>
The user wants to understand the dynamic changes ......
</PLANNING>
<CAPTION>
<BGM>Street noise; intermittent car horns.</BGM>
<SPEAKER>A: adult F; B: adult M; alternating turns.</SPEAKER>
<ASR></ASR>
<DESCRIPTION>The audio clip predominantly features static noise. ...... similar to that of a detuned television or a device failing to receive a signal.</DESCRIPTION>
</CAPTION>
<REASONING>
1. Identify changes in Intensity (Volume): ...... struggling to maintain a consistent output, adding to the impression of something malfunctioning or broken. </REASONING>
<SUMMARY>
The static noise in the audio is highly dynamic. ... leading to a sense of disorder and instability. </SUMMARY>
</THINK>
<RESPONSE>
The audio presents a static noise, ... is one of energetic chaos, preventing any possibility of calm or predictability.
</RESPONSE>

<REFLECT1> Does "A" mention the cake, not B? Check turn 3.</REFLECT1>
<NEW_RESPONSE>A</NEW_RESPONSE>
<REFLECT2> Does "A" mention the cake, not B? Check turn 3.</REFLECT2>
<NEW_RESPONSE>B</NEW_RESPONSE>

Prompt template of Refelection Sample

After producing the <RESPONSE>, you must perform a structured self-reflection step.
1. Compare the <RESPONSE> with the overall task requirements and check for issues such as: - Missing or incomplete coverage of the audio content (did it stop too early? were some speakers/segments missed?). - Repetition or redundant phrasing that should be removed or marked clearly. - Speaker attribution or diarization errors (wrong speaker assignment, merged speakers, or split speakers). - Prosody/tone/intonation mistakes or overemphasis on irrelevant details. - Inconsistent reasoning or labels (final choice must align with the reasoning and context). - Overly simplistic or single-hypothesis reasoning when alternatives exist.
2. Inside <REFLECT>...</REFLECT>, explicitly list: - The problems found in <RESPONSE>. - The corrections or adjustments needed (without referencing or leaking the gold standard answer text). - Any uncertainties or low-confidence areas.
3. Then rewrite the improved answer inside <FINAL_ANSWER>...</FINAL_ANSWER>, ensuring: - All necessary content is covered. - No hallucinated details are added beyond the given <CAPTION>, <ASR>, and <DESCRIPTION>. - Speaker attributions and reasoning are consistent. - The final answer matches the reasoning and is labeled correctly with confidence if required.
Format strictly as: <REFLECT> [Your structured reflection here] </REFLECT> <FINAL_ANSWER> [Your corrected, high-quality final answer here] </FINAL_ANSWER>
Here is the original bad answer: Turn0 Here is the golden answer: Golden_Ans

F Limitations of Simple ASR-Centric Text Reasoning

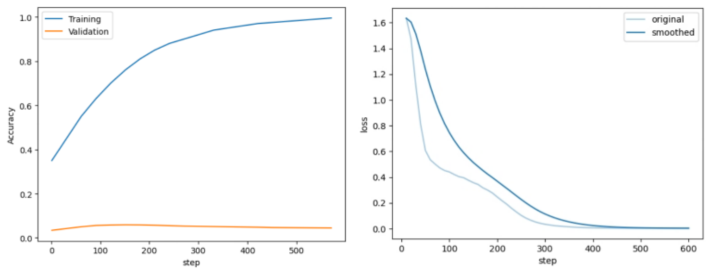
Early approaches to audio reasoning typically relied on converting speech into text via automatic speech recognition (ASR) and then performing reasoning over the textual transcript. While effective to some extent, this paradigm inevitably discards information that is uniquely embedded in the audio signal itself. To probe the limitations of this pipeline, we first evaluated the ASR+text reasoning approach on benchmarks such as CoVoST2 and MMAU. In CoVoST2, model performance is largely determined by raw ASR accuracy, and we observed that "simple ASR" signals are quickly memorized without yielding robust generalization. A case study is shown in Fig.**??**, which highlights several intrinsic challenges. Homophones and proper-name ambiguities necessitate long-range semantic modeling and external knowledge retrieval, while gendered pronouns in Chinese (e.g., "he/she") lack reliable acoustic cues and thus require contextual inference for disambiguation. In particular, Paraformer's frame-level alignment, coupled with strong language model priors, tends to induce a "nearest-neighbor copying" effect—yielding high accuracy on in-distribution transcripts but exhibiting pronounced failures under distributional shifts. Moreover, exposure to translation-oriented data (e.g., CoVoST2) can bias models such as Qwen-Audio to mistakenly trigger translation behavior, sometimes converting Chinese speech into other languages when acoustic cues are uncertain.

In Fig. 8(a), there is an improvement on base models if we asked them to answer questions with thinking in the format of `<THINK>...</THINK>` `<FINAL_ANSWER>...</FINAL_ANSWER>`. Therefore, we collected 2,050 samples from a subset of CoVoST2 (including 50 challenging cases reserved for the test set) and employed Kimi to generate CoT annotations. Using this data, we fine-tuned Qwen2-Audio and evaluated them on the designated test set. However, the models exhibited severe overfitting (see Fig. 7(b)) after only a single epoch of training: while the outputs consistently followed the required `<THINK>...</THINK>` `<FINAL_ANSWER>...</FINAL_ANSWER>` format and the training loss rapidly approached zero, the test accuracy dropped below 5%. This observation indicates that the gradients primarily optimized for surface-level grapheme mapping and fixed output formatting, without fostering genuine cross-sentence reasoning, coreference resolution, or knowledge-grounded inference.

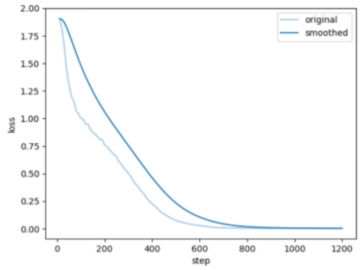
Consequently, these observations indicate that the "Thinking" component of chain-of-thought supervision should be allocated primarily to more challenging audio understanding tasks, such as multi-speaker dialogues and noisy environments—where reasoning signals genuinely drive the model to overcome semantic ambiguities and enforce knowledge-aware interpretations, rather than merely replicating templates on simple ASR tasks.

G The Use of Large Language Models (LLMs)

In order to reduce typos during the writing process and to optimize complex sentence structures so that the article becomes simpler and easier to read, we use mainstream large language models to refine certain paragraphs. For example, we use prompts such as "Help me correct the typos and grammatical errors in the above text, and streamline the logic to make it clear and easy to understand."

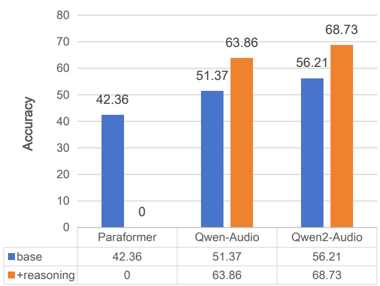


(a) Overfitting on Simple Translation Tasks of CoT fine-tuned model.

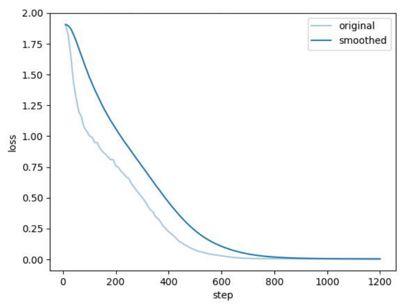


(b) Training loss on ASR of CoT fine-tuned model.

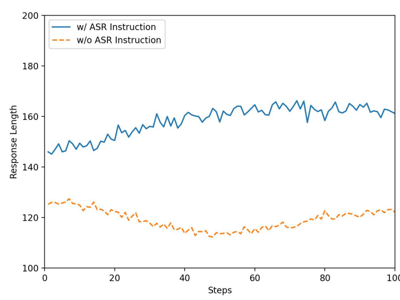
Figure 7: The training dynamics of a chain-of-thought (CoT) fine-tuned model (Qwen2-Audio-7B), indicating the model overfits to the training set in simple translation tasks. This suggests that CoT fine-tuning without additional regularization or more diverse data fails to yield robust generalization, particularly for tasks requiring broader reasoning beyond surface transcript matching.



(a) the comparison of base model and model with reasoning prompt.



(b) the training loss of fine-tuning Qwen2-Audio on our CoT-ASR data.



(c) the comparison of prompt with or without ASR finetuned on MSQA.

Figure 8: Experiments on the Exploration of Good Audio Reasoning prompt.