

THINK OUT LOUD, PAUSE IN SILENCE: CONFIDENCE-GUIDED REFLECT-PAUSE-ABORT FOR ROBUST AUDIO PERCEPTUAL UNDERSTANDING

Anonymous authors

Paper under double-blind review

ABSTRACT

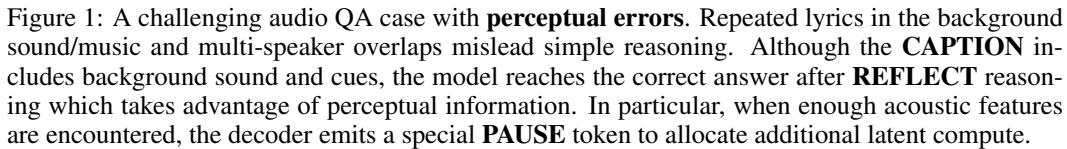
Large Audio Language Models (LALMs) mainly fail for two errors: perceptual errors misidentifying background sounds or speaker turns, and reasoning errors drifting rationales that decouple from acoustic evidence. To address these issues, we propose an adaptive framework that couples perceptual grounding with computation that expands only when needed. First, we introduce PAQA, a Perceptually grounded Audio QA dataset of 7,470 multiple-choice items that pairs multi-speaker, background-rich audio with stepwise reasoning and reflection annotations, enabling supervision of verifiable audio-grounded rationales. On the modeling side, we propose ConfAudio, which unifies explicit, reflective reasoning (fine-tuned on PAQA) with implicit, pause-driven latent computation trained via GRPO. A confidence-aware controller monitors lowest-group-confidence (LGC) during decoding to insert pauses when uncertainty rises and to abort unstable trajectories, thereby reallocating compute toward hard perceptual segments. To stabilize the training process, we design a composite reward that balances answer correctness, reasoning-answer consistency with perceptual robustness, and output format. Across PAQA, MMAU-mini, and MMAR, ConfAudio consistently improves both accuracy and consistency, particularly in noisy, multi-speaker conditions. Our results demonstrate that confidence-guided, adaptive reasoning—grounded in verifiable acoustic evidence—mitigates the dominant perceptual and reasoning failure modes in Audio-QA.

1 INTRODUCTION

Large language models (LLMs) have made notable progress in *reasoning* via chain-of-thought (CoT) prompting and reinforcement-learning (RL) post-training (Jaech et al., 2024; Guo et al., 2025), and similar advances have extended to visual modalities (Huang et al., 2025a; Feng et al., 2025). Unlike text, audio introduces unique challenges such as overlapping speakers, pronoun ambiguity, shifting emotions, and variable, noisy acoustic conditions. These factors often induce perceptual errors that are among the most prevalent failure modes of current models (Liu et al., 2024a).

Recent audio-capable LLMs (e.g., Qwen2-Audio (Chu et al., 2024), Audio Flamingo (Kong et al., 2024), SALMONN (Tang et al., 2023)) still tend to address audio question answering (Audio-QA) by mapping transcripts directly to answers, with limited verification against the underlying acoustic evidence. Prior audio CoT efforts (Zhang et al., 2024; Wang et al., 2024) supervise long free-form rationales but do not consistently yield improvements on challenging problems. Moreover, RL-only pipelines (Chen et al., 2024; Zhang et al., 2024; Xu et al., 2025; Zhou et al., 2025) improve answer accuracy, yet the explicit reasoning process itself has not shown consistent benefits for Audio-QA.

Previous work (Liu et al., 2024a) shows that dominant failures on the MMAR benchmark arise from perceptual errors and downstream reasoning mistakes. This underscores the need to first establish a strong perceptual foundation by explicitly incorporating verifiable evidence, especially in two high-frequency scenarios: (i) distinguishing environmental sounds, and (ii) accurately transcribing multi-speaker conditions. Importantly for audio reasoning, many acoustic cues (e.g., rhythmic density, timbre) cannot be faithfully translated into free-form text, so enforcing text-only rationales risks losing critical granularity. In real-world speech comprehension, humans often reflect on their reasoning process and, when uncertain, pause briefly to deliberate before responding. Inspired by



Interference from background sound and disambiguating multiple speakers are central challenges in real-world audio understanding, yet they remain underrepresented in existing datasets. To address this gap, we construct **PAQA**, the first **P**erceptually grounded **A**udio **Q**A dataset, coupling multi-speaker, background-rich audio with stepwise reasoning and reflection annotations. PAQA contains 7,470 multiple-choice Audio-QA pairs. Each example provides structured annotations—including background-music separation, speaker analysis, and multi-turn reflections—and records both an internal answer and a final response. As illustrated in Fig. 1, a model with simple reasoning incorrectly selects “Friday night” due to overreliance on repeated mentions, failing to distinguish background sound from actual conversation. In contrast, a model equipped with reflective reasoning isolates background sound and irrelevant speakers and aligns evidence to arrive at the correct answer.

(1) We construct **PAQA**, the first dataset that integrates multi-speaker and background-rich audio with stepwise reasoning and reflection annotations. It covers multi-speaker QA, noisy speech-to-text translation, and environment-centric QA, encouraging models to ground decisions in acoustic evidence rather than transcript shortcuts.

(3) We design a composite reward that jointly optimizes answer accuracy, format compliance, acoustic consistency, and length control, discouraging unnecessary verbosity. Experiments demonstrate that ConfAudio enhances adaptive reasoning, effectively reducing perceptual and reasoning errors.

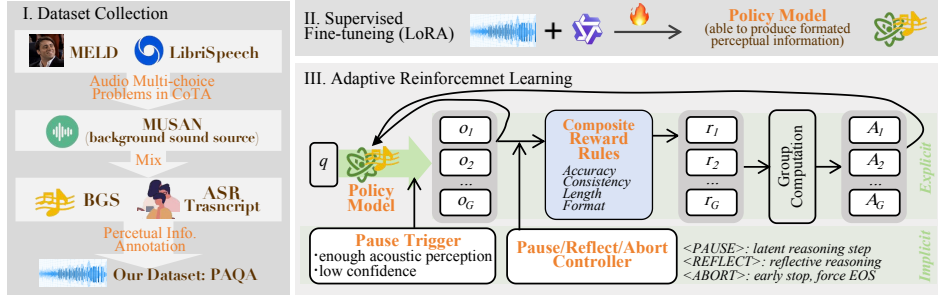


Figure 2: An overview of our work.

2 DATA COLLECTION FOR COMPLEX AUDIO UNDERSTANDING

2.1 COMPLEX AUDIO UNDERSTANDING

R1-AQA (Chen et al., 2024) and Omni-R1 (Zhou et al., 2025) show that requiring the model to read *write out* step by step text does not guarantee improvements in audio QA (AQA). In several AQA settings, explicit CoT provides only marginal or no gains over answer-only decoding, sometimes even increasing latency. For simple ASR or S2TT tasks, we also observed that models tend to overfit surface transcripts rather than perform robust reasoning over acoustic evidence (see Fig. 6). Unlike text-only scenarios, real-world audio understanding requires grounding in acoustic cues (e.g. speaker turns, overlapping speech regions) and careful attention to scene components. Motivated by this, we further analyze Qwen2-Audio’s bad cases on the CoTA (Zhang et al., 2024) benchmark and identify two major challenges:

- difficulties in handling multi-speaker conversations, where insufficient speaker diarization under overlapping or interleaved turns leads to incorrect attribution of utterances and a consequent loss of dialogue coherence;
- failures in environmental-sound discrimination, whereby non-speech events and nonsignificant background sound are classified as evidence;

To advance speaker-aware modeling and noise-resilient perception, we construct a dataset that integrates multi-speaker and background-rich audio with stepwise reasoning and reflection annotations.

Background-rich augmentation. We sample background audio from publicly licensed environmental categories in MUSAN (Snyder et al., 2015) (e.g. alarms, typing, rain, cafeteria, street traffic, soft instrumental music). For a clean speech clip s and a background clip n , we first RMS-normalize both and then scale the background so that the power ratio satisfies $\text{SNR}_{\text{dB}} = 10$, ensuring that the speech remains ten times stronger than the background, audibly present but not dominant. Each item is annotated with a tag indicating the presence and type of background (e.g., ‘Soft instrumental music – please ignore.’), which discourages unnecessary reliance on background cues.

Multi-speaker Alignment. To discourage models from shortcutting on global transcripts and to encourage speaker-localized reasoning, we annotate turn structures in a <SPEAKER> section using a compact, ordered format such as “**Speaker 1: ...**”. We then apply Qwen3-ASR (Team, 2025) to each audio sample to generate a verbatim raw transcript. To mitigate hallucination and drift between summaries and verbatim text, we introduce a quote-presence test (QPT), which measures fuzzy overlap between <ASR> snippets $A = a_i$ and <SPEAKER> sentences $S = s_i$. Specifically, SeqRatio is defined as the standard difflib ratio on normalized strings. Items with $QPT < 0.85$ are flagged for revision. The formulation is given by:

$$QPT = \frac{1}{M} \sum_{i=1}^M \max_{1 \leq j \leq N} \text{SeqRatio}(\text{norm}(s_i), \text{norm}(a_j)).$$

2.2 REFLECTION TO CORRECT WRONG INITIAL RESPONSES

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 the third column of Fig. 1. We then automatically detect errors—such as option mismatches, speaker attribution mistakes, hallucinated content inconsistent with ASR transcripts, or misinterpretation of noise cues—and prompt the model to produce a grounded diagnostic analysis `<REFLECT>`. 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 item, yields an additional corrected example, effectively doubling the supervised data while enriching them with interpretable, perception, grounded self-correction signals.

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 1.

Table 1: Dataset Source and Statistics.

Dataset Source	Main Skills Learning	BGM Used	Quantity	Reflection
Multi-Speaker (Zhang et al., 2024)	Multi-speaker Speech QA	Free Sound	1.5k	1.4k
MELD (Poria et al., 2019)	Speech Emotion QA	Sound Bible	1.5k	1.4k
CoVoST2 (Wang et al., 2020)	Speech-to-Text Translation	No	1.5k	No

3 METHODOLOGY

3.1 FRAMEWORK OVERVIEW

In real-world speech comprehension, humans often reflect on their reasoning process and, when uncertain, pause briefly to deliberate internally before responding (Kahneman (2011); Çalışkan & Narayanan (2020)). To mirror the sophisticated human ability to process complex audio environments through both explicit explanation and implicit contemplation (Goldman-Eisler (1968); Shriberg (2001)), we propose a **dual-mechanism, confidence-gated adaptive reasoning training pipeline**, with an internal confidence score determining the model’s reasoning path: an **explicit** Chain-of-Thought with `<REFLECT>`, or an **implicit** latent thinking to handle uncertainty with the `<PAUSE>` token. The whole framework is presented in Fig. 3.

Our methodology first employs **Supervised Fine-Tuning (SFT)** to teach the model a specific response structure, compelling it to generate answers that include both an analysis of perceptual audio information and a reflective reasoning process (with `<REFLECT>` in the chain-of-thought). Second, to move beyond simple pattern imitation and enhance the model’s adaptive reasoning capabilities, we utilize an advanced reinforcement learning algorithm, **Generalized Reward Policy Optimization (GRPO)**. Within this RL stage, we introduce a key innovation: when low confidence triggers the `<PAUSE>` mechanism, the model activates an **implicit latent reasoning module** to better process ambiguous, non-textual audio cues. This entire process is guided by a specialized reward function designed to optimize the quality and grounding of the final response.

3.2 SFT FOR STRUCTURED EXPLICIT GENERATION WITH REFLECTIVE REASONING

Using the specialized dataset detailed in Section 2.2, we explicitly train the model to output a dedicated `<REFLECT>` segment following its initial chain-of-thought reasoning. This structured output ensures that for each prompt, the model produces a transparent analysis of background sound and speaker, and reflection that allows for direct inspection before a final answer is given. 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.

Algorithm 1 Pseudocode of Pause/Abort/Reflect Mechaism in ConfAudio.

Require: query q , audio a ; policy model π_θ , reference model π_{ref} ; thresholds (τ_p, τ_a, τ_r) ; window K ; max pauses M ; max length L

```

1: function CONF( $\ell, \ell_{\text{ref}}, a$ ) ▷ entropy↓ + KL↓ + audio-uncertainty↓
2:   return clip( $w_1(1 - \text{H}(\text{softmax}(\ell))/\log |V|) + w_2(1 - \text{KL}(\ell||\ell_{\text{ref}})/\log |V|) + w_3(1 - u(a))$ ), 0, 1)
3: end function
4: function MELTDOWN( $y$ ) ▷ e.g., high  $n$ -gram repeat or abrupt entropy drop
5:   return (Repeat( $y$ ) >  $\gamma$ )
6: end function

7: function CONFAUDIOGENERATE( $q, a$ )
8:    $y \leftarrow []$ ,  $h \leftarrow []$ ,  $p \leftarrow 0$ ,  $low \leftarrow 0$ 
9:   for  $t = 1$  to  $L$  do
10:     $\ell \leftarrow \pi_\theta.\text{NEXTLOGITS}(q, a, y)$ ;  $\ell_{\text{ref}} \leftarrow \pi_{\text{ref}}.\text{NEXTLOGITS}(q, a, y)$ 
11:     $c \leftarrow \text{CONF}(\ell, \ell_{\text{ref}}, a)$ 
12:    if ( $c < \tau_a$ )  $\wedge$  ( $low \geq K \vee \text{MELTDOWN}(y)$ ) then
13:       $y.\text{APPEND}(<\text{EOS}>)$ ; break
14:    else if ( $c < \tau_p$ )  $\wedge$  ( $p < M$ ) then
15:       $y.\text{APPEND}(<\text{PAUSE}>)$ ;  $p \leftarrow p + 1$ 
16:       $h \leftarrow h \parallel \pi_\theta.\text{LATENTREASON}(q, a, y)$  ▷ generate until </PAUSE>
17:      continue
18:    else
19:       $tkn \leftarrow \text{SAMPLE}(\ell)$ ;  $y.\text{APPEND}(tkn)$ 
20:       $low \leftarrow (c < \tau_p) ? (low + 1) : 0$ 
21:      if  $tkn = <\text{EOS}>$  then break
22:    end if
23:  end if
24:  end for
25:   $c_g \leftarrow \text{AGGCONF}(y)$ 
26:   $need \leftarrow (c_g < \tau_r) \vee \neg \text{CONSISTENT}(q, y)$ 
27:  if  $need$  then
28:     $y_{\text{pre}} \leftarrow y$ ;  $(y_{\text{ref}}, y) \leftarrow \pi_\theta.\text{REFLECTANDREVISE}(q, a, y_{\text{pre}})$ 
29:  else
30:     $y_{\text{ref}} \leftarrow \emptyset$ ,  $y_{\text{pre}} \leftarrow \emptyset$ 
31:  end if
32:  return  $y$ ,  $\{y_{\text{pre}}, y_{\text{ref}}, h, \text{pauses} = p, \text{aborted} = (y[-1] = <\text{EOS}>)\}$ 
33: end function

```

3.3 RL WITH ADAPTIVE REASONING

3.3.1 GRPO ON FINE-TUNED MODELS

We observed that fine-tuning instruction-tuned models on PAQA leads to performance improvements, although these gains are not significant (see Table 2). While the Supervised Fine-Tuning (SFT) stage effectively establishes a foundation by teaching the model to generate more complex chain-of-thought reasoning, its potential may be inherently constrained by the nature of imitation learning. To transcend these limitations, we introduce a second stage based on Group-aware Relative Policy Optimization (GRPO).

Starting from the SFT checkpoint (reference policy π_{ref} frozen), we generate groupwise rollouts, compute $R(\mathbf{z})$ via equation 4, and update π_θ with GRPO. We partition rollouts by task group $g \in \{\text{PAQA}, \text{AVQA}\}$ and difficulty bucket (low/med/high). Within each group, we compute groupwise baselines to reduce variance:

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

Let $A^{(i)} = \tilde{R}^{(i)}$ be the advantage. We optimize a clipped PPO objective over tokens t :

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E} \left[\min \left(\rho_t^{(i)} A^{(i)}, \text{clip}(\rho_t^{(i)}, 1 - \epsilon, 1 + \epsilon) A^{(i)} \right) \right], \quad \rho_t^{(i)} = \frac{\pi_\theta(z_t^{(i)} | \cdot)}{\pi_{\theta_{\text{old}}}(z_t^{(i)} | \cdot)}.$$

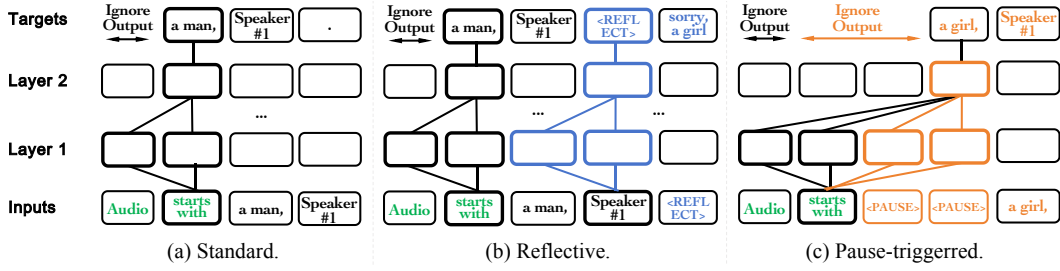


Figure 3: Standard vs. reflective-inference vs. pause-inference (and finetuning). Similar to Goyal et al. (2024), the rounded squares denote one Transformer operation (a self-attention and MLP) in a 2-layer Transformer. Any Ignore Output denotes that during inference, the corresponding output token is not extracted and thus, not fed back autoregressively.

To process and reason upon non-textual audio cues that are often lost in pure text representations, we incorporate confidence into GRPO by using the lowest-group-confidence (LGC) as a sample weight (see in 3.3.2). For a trajectory i with task reward r_i^{task} (covering correctness, formatting, consistency, and length penalties) and group baseline \bar{r} , the advantage is calculated as

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

where w_i is a clipped, standardized function of the trajectory’s LGC, and $w_i = 0$ for filtered samples. The final reward may also include recovery and leak terms when pause control is enabled:

$$r_i = r_i^{\text{task}} + \eta \cdot \max(0, \text{LGC}^{\text{post}} - \text{LGC}^{\text{pre}}) - \lambda_{\text{leak}} \cdot \mathbf{1}\{\text{leak}\},$$

where $\eta, \lambda_{\text{leak}}$ are selected to balance accuracy, formatting, and robustness.

3.3.2 IMPLICIT LATENT THINKING

Beyond explicit reflection, the model can also perform hidden computation via implicit pauses. This mechanism is particularly vital when processing complex audio data. Audio streams contain a wealth of perceptual, non-verbal cues—such as speaker intonation, overlapping speech, or ambient noise—that are often difficult to articulate fully in explicit text. An implicit pause allows the model to internally process and weigh these subtle audio features before committing to a textual output, enabling a deeper, more grounded reasoning process.

To govern this behavior, we introduce a confidence-based gating mechanism with two distinct thresholds: τ_{pause} for outright termination and a softer threshold $\tau_{\text{pause}} > \tau_{\text{abort}}$. At each decoding step t , a specific action is selected based on the model’s internal confidence score $C_{k(t)}$:

$$\text{action}(t) = \begin{cases} \text{abort} & \text{if } C_{k(t)} < \tau_{\text{abort}}, \\ \text{pause} & \text{if } \tau_{\text{abort}} \leq C_{k(t)} < \tau_{\text{pause}}, \\ \text{continue} & \text{otherwise.} \end{cases} \quad (1)$$

where `continue` is the default action to normal generation. When confidence falls into the intermediate range, the model triggers an internal deliberation step. When a pause is triggered at t^* , the model emits a `<PAUSE>` special token and generates up to K_p latent tokens $\hat{\mathbf{z}}_{1:L}$ (where $L \leq K_p$). These “pause tokens” are *not* surfaced in the final visible output and are explicitly excluded from the gradient calculations during parameter optimization. Their sole function is to iteratively update the model’s internal hidden state, allowing it to “think” and process complex information before resuming the generation of visible tokens. Formally, the full token sequence and visible output are defined as (\oplus means concatenation):

$$\tilde{\mathbf{y}} = \mathbf{y}_{1:t^*} \oplus \text{<PAUSE>} \oplus \hat{\mathbf{z}}_{1:L}, \quad \mathbf{y}_{\text{vis}} = \mathbf{y}_{1:t^*}. \quad (2)$$

To encourage pausing specifically in text involving speech-only, non-textual cues (where extra acoustic reasoning may be beneficial), we maintain a keyword set \mathcal{A} , such as “tone”, “pitch”, whenever any $w \in \mathcal{A}$ appears in the recent context, we add a positive logit bias $\beta_{\text{ac}} > 0$ to the `<PAUSE>` token:

$$\ell_{\text{<PAUSE>}} \leftarrow \ell_{\text{<PAUSE>}} + \beta_{\text{ac}} \cdot \mathbb{I}[\exists w \in \mathcal{A} \text{ in context}]. \quad (3)$$

Crucially, the same confidence signal $\{C_k\}$ that defines $\text{LGC}(\mathbf{y}) = \min_k C_k$ during training also governs the runtime policy above (pause vs. continue vs. abort). This shared mechanism ties sample selection (via top- p filtering on LGC) to inference-time latent compute allocation (via `<PAUSE>` triggers) and safety (via abort), yielding a coherent control scheme.

Furthermore, `abort` action serves as a crucial efficiency mechanism. If the confidence score is extremely low ($C_{k(t)} < \tau_{\text{abort}}$), it indicates a high likelihood that further deliberation will be unproductive. The model is therefore instructed to terminate the current generation path immediately, for example, by forcing the output of an end-of-sequence (`<EOS>`) token. This strategy prevents the model from getting stuck in prolonged, unnecessary pause loops and significantly accelerates overall inference speed by pruning unpromising trajectories early.

3.3.3 LOWEST GROUP CONFIDENCE (LGC)

We consider the lowest group confidence metric to provide sufficient signals for estimating the quality of reasoning trajectory by identifying its most uncertain segment. Given an input x and a sequence of completion tokens $\mathbf{y} = (y_1, \dots, y_T)$ with per-token log-probabilities $\ell_t = \log P_\theta(y_t | x, y_{<t})$. The sequence is first partitioned into K overlapping windows $\mathcal{W}k$ of length w with stride s . For each window, we compute a normalized geometric mean probability:

$$C_k = \exp\left(\frac{\sum_{t \in \mathcal{W}k} m_t \ell_t}{\sum_{t \in \mathcal{W}k} m_t + \delta}\right),$$

where $m_t \in \{0, 1\}$ masks out prompt and padding tokens, and δ avoids division by zero. The LGC of the trajectory is then defined as the minimum of these window confidence scores:

$$\text{LGC}(\mathbf{y}) = \min_{k=1, \dots, K} C_k$$

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 Wang et al. (2025).

3.4 COMPOSITE REWARD FUNCTION

To produce correct answers but also generate interpretable and well-organized reasoning, we introduce a novel composite reward function R defined as the weighted sum of four components: besides the accuracy reward and a format reward, we designed a consistency reward for reasoning coherence and explanation quality, and length reward when accurate for stabilize the training.

$$R = w_{\text{acc}} \mathcal{R}_{\text{acc}} + w_{\text{fmt}} \mathcal{R}_{\text{fmt}} + w_{\text{cons}} \mathcal{R}_{\text{cons}}(\hat{y}, \hat{y}_{\text{CoT}}) + w_{\text{len}} (\mathcal{R}_{\text{acc}} \times \mathcal{R}_{\text{len}}), \quad (4)$$

where \mathcal{R}_{acc} measures the correctness of the final answer, the format reward, denoted as \mathcal{R}_{fmt} , validates whether the output follows the required structured reasoning schema, $\mathcal{R}_{\text{cons}}(\hat{y}, \hat{y}_{\text{CoT}})$ measures the logical consistency between the final answer and the intermediate reasoning, and \mathcal{R}_{len} penalizes both under- and over-elaboration.

3.4.1 ACCURACY REWARD.

The accuracy reward, \mathcal{R}_{acc} , measures the correctness of the final answer. We parse the completion x for the `<FINAL_ANSWER>` (FA) tag, falling back to the `<RESPONSE>` (R) tag if FA is absent. From this, we extract the predicted answer $\hat{y}_L(x) \in \{A, B, C, D\}$ and compare it to the ground truth label. Let a completion be denoted by x . We parse both `<FINAL_ANSWER>` (FA) and `<RESPONSE>` (R), taking FA as the primary channel and falling back to R when FA is absent. From x , we extract a predicted letter and a channel text $\hat{y}_T(x)$, defined as the raw inner text of FA or R. Within `<THINK>`, let $\mathcal{A} = a_j$ denote ASR sentences and $\mathcal{S} = s_i$ denote speaker snippets (i.e., quoted spans or colon-delimited clauses).

3.4.2 FORMAT REWARD.

The Format Reward validates whether the output follows the required structured reasoning schema; it is assigned 1.0 if all required tags are present, correctly ordered, and properly nested, and 0.0

otherwise. This reward component addresses a common challenge in Reinforcement Learning (RL) known as reward sparsity. We distinguish between two levels of formatting:

Strict Format (Ideal Schema): This is the ideal, fully-specified structure. It requires the completion to follow a complex schema. However, enforcing this strict format during RL training is problematic. The probability of a model, especially early in training, generating this exact complex structure by chance is near zero. This would result in a persistent $\mathcal{R}_{\text{fmt}} = 0.0$, providing no effective learning signal (gradient) for the model to learn the format.

Weak Format (RL Reward for Latent Reasoning): To overcome reward sparsity, \mathcal{R}_{fmt} is implemented based on a "weak format". This "weak format" approach allows the model to first learn the fundamental `<think>...<response>` sequence. Subsequently, other reward components, such as the consistency reward, implicitly incentivize the model to correctly populate the CoT with the necessary analysis (e.g., ASR, Speaker analysis) to maximize its score, thereby progressively guiding the model toward the ideal "strict format".

3.4.3 CONSISTENCY REWARD.

Beyond task accuracy and format concerns, we regularize chains for *internal consistency* along three axes: (i) **BGM robustness** blocks spurious causal use of background sound/music; (ii) **Speaker-ASR fidelity** rewards quotes/snippets that appear in the ASR transcript; and (iii) **Reasoning-Answer consistency** rewards agreement between the last internal choice from `<THINK>` and the final answer. Let $r_{\text{bgs}}(x) \in \{0, 1\}$, $r_{\text{spk}}(x) \in [0, 1]$, and $r_{\text{ra}}(\hat{y}, \hat{y}_{\text{CoT}}) = \mathbf{1}[\tilde{y}_L(\hat{y}, \hat{y}_{\text{CoT}}) = \hat{y}_L(\hat{y}, \hat{y}_{\text{CoT}})] \in \{0, 1\}$. We combine them with a hard background sound/music gate and a convex mixture:

$$\mathcal{R}_{\text{cons}}(\hat{y}, \hat{y}_{\text{CoT}}) = r_{\text{bgs}}(\hat{y}, \hat{y}_{\text{CoT}}) \left(\lambda_{\text{spk}} r_{\text{spk}}(\hat{y}, \hat{y}_{\text{CoT}}) + \lambda_{\text{ra}} r_{\text{ra}}(\hat{y}, \hat{y}_{\text{CoT}}) \right), \quad \lambda_{\text{spk}}, \lambda_{\text{ra}} \geq 0, \lambda_{\text{spk}} + \lambda_{\text{ra}} = 1. \quad (5)$$

In words, any offending background sound/music causal claim zeroes the consistency reward; otherwise, we interpolate between sentence-level Speaker-ASR alignment and self-agreement of the final answer. The default weights $\lambda_{\text{spk}} = \lambda_{\text{ra}} = 0.5$ worked well in our runs.

Background sound/music robustness. To prevent spurious cues from background sound, we penalize any reasoning/description sentence that *uses background sound/music as causal evidence*. We set $r_{\text{bgs}}(x) = 0$ if reasoning invokes background sound/music as causal evidence (e.g., "bgm", "background sound", instrument names), otherwise $r_{\text{bgs}}(x) = 1$.

Speaker-ASR fidelity We measure how faithfully speaker-attributed quotes in the model's outputs appear in the ASR transcript. Let $\mathcal{S} = \mathcal{S}(\hat{y}, \hat{y}_{\text{CoT}})$ be the set of quoted/snippet spans extracted from the chain (e.g., speaker-attributed quotes), and let \mathcal{A} be the set of sentences from the ASR transcript of the input audio. We softly align each snippet to its best-matching ASR sentence via a normalized similarity $\text{sim}(\cdot, \cdot) \in [0, 1]$, and average over snippets:

$$r_{\text{spk}}(\hat{y}, \hat{y}_{\text{CoT}}) = \frac{1}{|\mathcal{S}|} \sum s \in \mathcal{S} \max_{a \in \mathcal{A}} \text{sim}(\text{norm}(s), \text{norm}(a)) \in [0, 1]. \quad (6)$$

Here, $\text{norm}(\cdot)$ is a deterministic text normalization (lowercasing, stripping punctuation/speaker tags, collapsing whitespace). The similarity is a normalized edit-based score:

$$\text{sim}(u, v) = 1 - \frac{\text{Lev}(u, v)}{\max\{|u|, |v|\}} \in [0, 1], \quad (7)$$

where $\text{Lev}(\cdot, \cdot)$ is the (character-level) Levenshtein distance.

This rewards quotes that truly occur in the ASR while tolerating minor lexical or punctuation variations; fabricated or hallucinated quotes receive low similarity. For completeness, we adopt the following conventions for edge cases:

$$r_{\text{spk}}(\hat{y}, \hat{y}_{\text{CoT}}) = \begin{cases} 1, & |\mathcal{S}| = 0 \text{ (no verbatim quotes to verify);} \\ 0, & |\mathcal{A}| = 0 \wedge |\mathcal{S}| > 0 \text{ (no ASR to support quotes);} \end{cases} \quad (8)$$

Reasoning–Answer consistency. In long, noisy chains, decoding drift can make the model “reason to A but output C ”. Therefore, we design rewarding self-agreement. From `<THINK>` we extract the last declared option $\tilde{y}_L(x)$. Let $\hat{y}_L(x)$ denote the final emitted letter (FA/R). Then we calculate $r_{\text{cons}}(x) = \mathbf{1}[\tilde{y}_L(x) = \hat{y}_L(x)]$.

Together, the consistency measure discourages spurious reliance on background sound/music, enforces alignment between speaker summaries and ASR evidence, and ensures that the internally declared answer matches the final emitted answer.

3.4.4 LENGTH SHAPING WHEN ACCURATE.

To encourage adequate evidence gathering (speaker attribution, noise filtering) without overlong chains, we introduce a length sub-reward when the accuracy reward is 1. Let $T(x)$ be a simple token proxy (count of non-whitespace sequences in the *whole* completion). With thresholds $T_{\min} = 300$, $T_{\max} = 600$, and decay scale $K > 0$, we use a piecewise-linear schedule:

$$\mathcal{R}_{\text{len}} = \begin{cases} \frac{T(x)}{T_{\min}}, & 0 \leq T(x) < T_{\min}, \\ 1, & T_{\min} \leq T(x) \leq T_{\max}, \\ \max(0, 1 - \frac{T(x) - T_{\max}}{K}), & T(x) > T_{\max}. \end{cases} \quad (9)$$

To enforce clean outputs, we gate by the absence of post-answer content, where any non-whitespace after `</FINALANSWER>` leads to 0.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

All experiments fine-tune the same pretrained backbones (Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B), using the framework introduced by Chen et al. (2024). 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 latent thinking mechanism, we set `tau_pause_quantile=0.50` that allows up to 8 pauses per sequence with 64 thinking tokens each, plus recovery bonus (0.05) and leak penalty (1.0) for think token containment. We evaluate six configurations: **SFT**, supervised fine-tuning only with no reasoning schema and no `<pause>`; **GRPO-NoThink**, GRPO post-training that emits answers directly 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 (ConfAudio)**, GRPO enhanced with the explicit schema and `<pause>`; and **External Baselines** including Audio-Reasoner (Zhang et al., 2024) and Audio-Thinker (Xu et al., 2025). We use PAQA (8k question-answering pairs) for supervised finetuning; and leverage 30,000 data from AQVA (Yang et al., 2022) for the RL part. Models are evaluated on several benchmarks **PAQA Test(hard)**, **MMAU** (Park et al., 2024), and **MMAR** (Liu et al., 2024a), the results are listed below.

4.2 MAIN RESULTS

Table 2 summarizes the main results across two aggregated benchmarks (MMAU, MMAR). Several key findings emerge from the experimental analysis.

SFT. vs. Baseline Injecting audio perceptual information via SFT proves beneficial, particularly in speech, which constitute the primary focus of our study. However, as the data is heavily skewed toward speech, some degree of domain shift is observed in music, leading to suboptimal results.

RL vs. SFT *GRPO-NoThink* performs better than with SFT, and each introduced module in our method improves the performance well. This demonstrates that optimizing structured reasoning and perceptual-consistency signals is essential to enhance audio perception reasoning.

Pause mechanism works. Ours method with explicit reflection and implicit latent reasoning surpasses all other baselines. Notably, it improves MMAU-Music by +3.29 and MMAR-Music by

Table 2: Performance on MMAU Test-mini (Park et al., 2024) and MMAR (Liu et al., 2024a).

Method	MMAU Test-mini \uparrow				MMAR \uparrow			
	Sound	Music	Speech	Average	Sound	Music	Speech	Average
Qwen2-Audio	61.26	53.59	48.05	54.30	33.33	24.27	32.31	30.00
+SFT	62.76	44.61	55.86	54.41	41.82	34.95	45.92	40.90
+GRPO-NoThink	68.17	61.38	60.66	63.40	51.52	38.83	45.92	45.40
+GRPO+CoT	70.27	59.88	59.46	63.20	58.18	33.98	46.60	46.30
+GRPO+ExpCoT	75.07	58.98	63.66	65.90	44.85	39.81	59.86	48.20
& - weak format	72.97	61.08	63.96	66.00	42.42	43.69	61.22	49.10
Ours (ConfAudio)	75.67	62.27	64.26	67.40	58.18	45.63	62.59	55.50
Audio-CoT	62.16	55.99	56.16	58.10	35.76	25.24	34.01	31.67
Audio-Reasoner	60.06	64.30	60.70	61.71	43.64	33.50	32.99	36.71

+5.82, indicating that the pause latent thinking mechanism is particularly beneficial in complex acoustic settings. More detailed analysis are provided in Appendix.

Abort mechanism balances speed with formatting penalties. While abort (early stopping under high uncertainty/high entropy) improves throughput and latency control, we observe that it tends to output think-only simplified content, yielding very low format rewards. We therefore down-weight the format reward and make a baseline with a weak format. Even under this setting, ConfAudio continues to outperform alternatives, suggesting that the pause mechanism remains the principal driver of performance gains.

4.3 ABLATION STUDY

Background Sound As shown in Fig.4(a), the introduction of background sound leads to measurable degrades zero-shot performance. However, this drop is substantially mitigated while explicit “ignore prompt” cues are provided. This validates that our reflection step (Section 3) substantially improve accuracy. In Fig. 4(b), we further compare the effect of varying numbers of reflection turns, moving from 0 to 1 round yields a large accuracy enhancement while keeping outputs mostly clean. However, adding more rounds brings diminishing returns and “overthinking”. A similar trade-off is identified in the use of the pause mechanism. Excessive pausing negatively affects performance(see Fig. 5), suggesting that it is suitable to set max pause token between 1 and 3. We also evaluate on the test set of PAQA(see Tab.3), on the category of multi-speaker and MELD (Zhang et al., 2024), ConfAudio performs the best. Furthermore, under the challenging setting with background sound at SNR=5dB, a condition that considerably degrades most models, our ConfAudio 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.

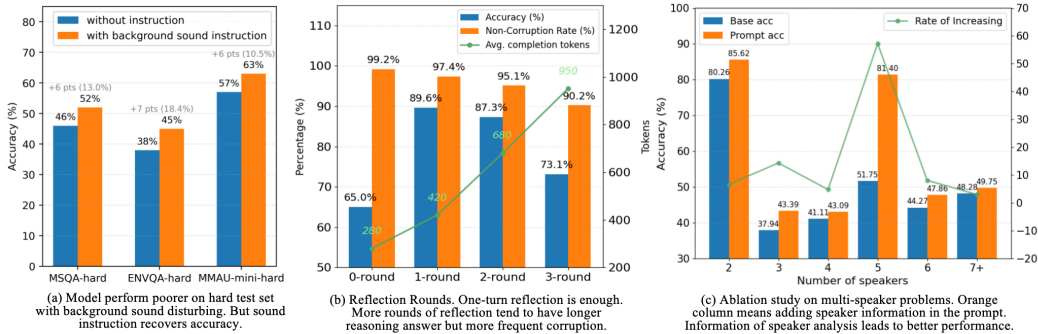


Figure 4: Comparison between different audio situations.

Multi Speaker Overall, adding a short “background information” prompt improves accuracy from 49.95 \rightarrow 58.16 on 3,071 items (+8.21 pts, +16.4% relative). The benefit is consistent across all speaker counts, confirming that context priming helps the model ground references and disambiguate turns. Difficulty scales with crowding: the base model is strong with 2 speakers (80.26%), but drops sharply as speakers increase. This pattern matches the intuition that more speakers introduce

Model	Multi-Speaker(hard)		BGM-rich Acc.	
	Acc.	Consistency \uparrow	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

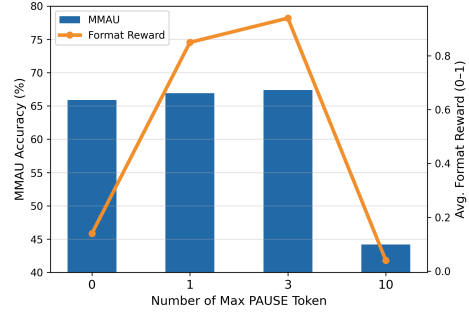


Table 3: Evaluation on the test set of PAQA. Our model performs best in each category.

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

attribution and coreference errors. Crucially, 3–6 speaker cases benefit most from the added context. For 7+ speakers, the improvement is modest, indicating that a simple speaker hint is insufficient once turn-taking becomes very dense; richer cues (explicit diarization tags, role summaries, or brief scene summaries) are likely needed.

Length Reward We also observe some findings about the design of length-reward. 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 (See Fig. 8).

5 RELATED WORKS

5.1 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., 2023) advanced ASR, but remained fragile in real-world reasoning tasks involving overlapping speakers and non-stationary noise. On-demand CoT in Audio Flamingo 3(Chen et al., 2025) and structured CoT in Audio-Reasoner (Zhang et al., 2024)—yet models often reverted to transcript shortcuts whenever acoustic evidence was difficult to verbalize. 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 purely textual evidence.

5.2 REASONING IN (AUDIO-)LANGUAGE MODELS: EXPLICIT AND REFLECTIVE

In LLMs, structured reasoning through CoT, reflection, and RL post-training has yielded consistent gains beyond supervised fine-tuning (SFT) (Guo et al., 2025; Kimi Team, 2025). Replication, robustness, and efficiency studies further refined these pipelines, while Vision-R1 and Video-R1 extended RL-based reasoning to perceptual tasks with overthinking suppression (Yu et al., 2025; Huang et al., 2025a;b; Feng et al., 2025). In audio, GRPO-style RL underlies R1-AQA and Omni-R1 (Shao et al., 2024; Chen et al., 2024; Zhou et al., 2025), with mixed evidence on whether RL alone suffices. More recent approaches (Liu et al., 2024b; Xu et al., 2025; Li et al., 2025; 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, grounding, and conciseness.

5.3 IMPLICIT REASONING AND PAUSE-GATED LATENT COMPUTE

Complementary to explicit rationales, implicit computation allocates additional internal processing before token emission. Learned <pause> tokens can trigger silent forward passes (Goyal et al., 2023), echoing earlier adaptive-computation approaches(Graves, 2016; 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, ConfAudio diverts into a short, budgeted latent stream and can abort tail trajectories under severe uncertainty.

6 CONCLUSION

In this paper, to address two entangled failure modes in Audio-QA, perceptual and reasoning errors, we build **PAQA** (7,470 items) to supervise verifiable, audio-grounded rationales, and propose **ConfAudio**, which couples explicit reflection with implicit, pause-driven latent thinking trained via GRPO with a composite reward. Specially, a lowest-group-confidence controller inserts `<pause>` or aborts unstable trajectories. ConfAudio delivers consistent gains in accuracy and consistency under noisy, multi-speaker conditions, narrowing the gap between acoustic evidence and reasoning.

ETHICAL CONSIDERATIONS.

Our dataset is constructed from publicly available corpora or controlled augmentations, with all speech either anonymized or synthesized to avoid privacy leakage. Despite the contributions, several limitations remain. First, while our dataset is carefully annotated with multi-speaker and background-rich reasoning structures, its scale is modest compared to general-purpose audio corpora, which may limit coverage of rare conversational phenomena.

REPRODUCTIVITY STATEMENT.

We prioritize reproducibility by releasing dataset specifications, and preprocessing scripts for background injection, speaker segmentation, and ASR alignment. All training configurations—including optimizer settings, batch sizes, learning rate schedules, and LoRA ranks—are documented and released as YAML files. Our evaluation follows a consistent protocol across our dataset, MMAU, and MMAR, reporting accuracy, consistency, and robustness under noise. Results are averaged over multiple random seeds to avoid cherry-picking. Upon publication, we will release our training data, code, inference pipelines, and checkpoints under an open-source license.

REFERENCES

- Andrea Banino, Samuel Ritter, et al. Pondernet: Learning to ponder. In *ICML*, 2021.
- A Çalıskan and S Narayanan. Analyzing cognitive load from speech: A survey. *IEEE Signal Processing Magazine*, 37(5):72–85, 2020.
- Bo Chen, Yujia Peng, Rui Chen, Ziyang Ma, and Yuexian Zou. R1-AQA: Reinforcement learning with GRPO for audio question answering. *arXiv preprint arXiv:2409.11371*, 2024. URL <https://arxiv.org/abs/2409.11371>.
- Rui Chen, Yujia Peng, Ziyang Ma, Zhenyang Ni, Chen Zhang, and Yuexian Zou. Audio flamingo 3: On-demand chain-of-thought for audio reasoning. *arXiv preprint arXiv:2502.08352*, 2025. URL <https://arxiv.org/abs/2502.08352>.
- 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. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.
- Kaituo Feng et al. Video-r1: Reinforcing video reasoning in mllms. 2025. URL <https://arxiv.org/abs/2503.21776>.
- F Goldman-Eisler. *Psycholinguistics: Experiments in spontaneous speech*. Academic Press, 1968.
- Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens, 2024.

- Sachin Goyal et al. Think before you speak: Training language models with pause tokens. 2023. URL <https://arxiv.org/abs/2310.02226>.
- Alex Graves. Adaptive computation time for recurrent neural networks. In *ICML*, 2016.
- Dongliang Guo, DeepSeek-AI, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. 2025. URL <https://arxiv.org/abs/2501.12948>.
- Wenxuan Huang et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. 2025a. URL <https://arxiv.org/abs/2503.06749>.
- Yuchen Huang, Yibo Wang, Shaoguang Mao, Wenshan Wu, Qi Liu, Enhong Chen, and Furu Wei. Progressive thinking suppression training for mitigating overthinking. *arXiv preprint arXiv:2502.10351*, 2025b. URL <https://arxiv.org/abs/2502.10351>.
- Aaron Jaech et al. Openai-o1: Learning to reason with reinforcement learning from verifiers. 2024. URL <https://arxiv.org/abs/2412.16720>.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O. Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning, 2025. URL <https://arxiv.org/abs/2503.09516>. *arXiv preprint arXiv:2503.09516*.
- Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
- Kimi Team. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL <https://arxiv.org/abs/2501.12599>.
- Zhifeng Kong et al. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. 2024. URL <https://arxiv.org/abs/2402.01831>.
- Guangyao Li et al. Reinforcement learning outperforms supervised fine-tuning for audio question answering. 2025. URL <https://arxiv.org/abs/2503.11197>.
- Chen Liu, Yujia Peng, Rui Chen, Ziyang Ma, and Yuexian Zou. MMAR: Multi-modal audio reasoning benchmark. *arXiv preprint arXiv:2505.13032*, 2024a. URL <https://arxiv.org/abs/2505.13032>.
- Yixuan Liu, Yujia Peng, Rui Chen, Ziyang Ma, and Yuexian Zou. SARI: Structured audio reasoning with curriculum and reinforcement learning. *arXiv preprint arXiv:2409.11372*, 2024b. URL <https://arxiv.org/abs/2409.11372>.
- Aman Madaan et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL <https://arxiv.org/abs/2303.17651>.
- Jae Park, Yujia Peng, Rui Chen, Ziyang Ma, and Yuexian Zou. MMAU: A benchmark for multimodal audio understanding and reasoning. *arXiv preprint arXiv:2406.06318*, 2024. URL <https://arxiv.org/abs/2406.06318>.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019. Association for Computational Linguistics.
- Zhen Shao et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. 2024. Introduces Group Relative Policy Optimization (GRPO).
- Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL <https://arxiv.org/abs/2303.11366>.
- Elizabeth Shriberg. To 'err' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169, 2001.

- David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus, 2015.
- Yue Tang, Hongyu Lan, Guangzhi Sun, Xianjun Xia, Mengyue Wu, Yuping Wang, Jun Zhang, Zejun Ma, and Yuexian Zou. SALMONN: Speech-audio-language modeling for open-ended understanding. *arXiv preprint arXiv:2307.00162*, 2023. URL <https://arxiv.org/abs/2307.00162>.
- Qwen Team. Qwen3 technical report. Technical report, Qwen Team, 2025. URL https://github.com/QwenLM/Qwen3/blob/main/Qwen3_Technical_Report.pdf. Accessed: 2025-09-25.
- Changhan Wang, Anne Wu, and Juan Pino. Covost 2: A massively multilingual speech-to-text translation corpus, 2020.
- Hao Wang, Yujia Peng, Rui Chen, Ziyang Ma, and Yuexian Zou. Audio-CoT: Chain-of-thought supervision for audio-language models. *arXiv preprint arXiv:2401.13969*, 2024. URL <https://arxiv.org/abs/2401.13969>.
- Xuezhi Wang et al. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL <https://arxiv.org/abs/2203.11171>.
- Yifei Wang, Yutong Wang, Zhengyang Zhou, Yifan Zhang, Zhengjue Wang, Zhiheng Xi, Chenjun Xiao, and Yang Yuan. Deep think with confidence: Uncertainty-aware reasoning in llms via scaling verification, 2025.
- Tian Xu, Yujia Peng, Rui Chen, Ziyang Ma, and Yuexian Zou. Audio-thinker: Adaptive reflective reasoning for audio-language models. *arXiv preprint arXiv:2502.14457*, 2025. URL <https://arxiv.org/abs/2502.14457>.
- Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3480–3491, 2022.
- Wenhao Yu, Qi Zhu, Chuang Niu, Sharath Chandra Raparthy, Kristian Greenewald, Hao Wang, Yoon Kim, and Tommi Jaakkola. Efficient reinforcement learning for long-chain reasoning. *arXiv preprint arXiv:2502.12345*, 2025. URL <https://arxiv.org/abs/2502.12345>.
- Li Zhang, Yujia Peng, Rui Chen, Ziyang Ma, and Yuexian Zou. Audio-reasoner: Large-scale structured CoT for audio reasoning. *arXiv preprint arXiv:2406.06317*, 2024. URL <https://arxiv.org/abs/2406.06317>.
- Lin Zhou, Yujia Peng, Rui Chen, Ziyang Ma, and Yuexian Zou. Omni-R1: GRPO fine-tuning of qwen2.5-omni for audio QA. *arXiv preprint arXiv:2504.12207*, 2025. URL <https://arxiv.org/abs/2504.12207>.

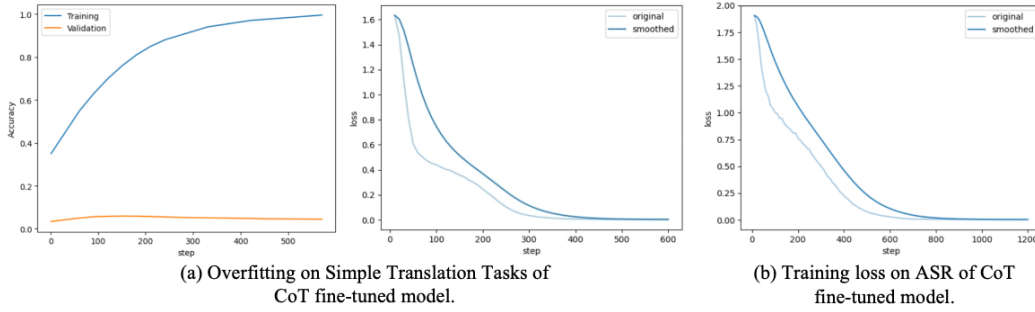


Figure 6: The training dynamics of a chain-of-thought (CoT) fine-tuned model (Qwen2-Audio-7B), indicating the model overfits to the training set in simple translation tasks. This suggests that CoT fine-tuning without additional regularization or more diverse data fails to yield robust generalization, particularly for tasks requiring broader reasoning beyond surface transcript matching.

A DATA COLLECTION AND EXPLORATION OF HARD AUDIO UNDERSTANDING

A.1 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.9, 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.

As shown in Fig.7(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 chain-of-thought style annotations. Using this data, we fine-tuned Qwen2-Audio and Qwen-Audio and evaluated them on the designated test set. However, the models exhibited severe overfitting(see Fig.6(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.

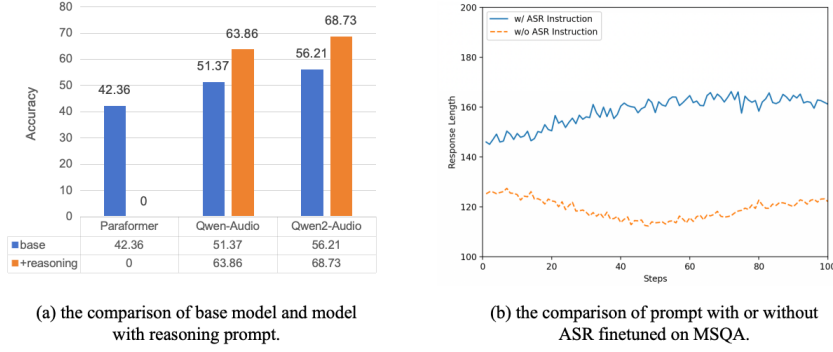


Figure 7: Experiments on the Exploration of Good Audio Reasoning prompt.

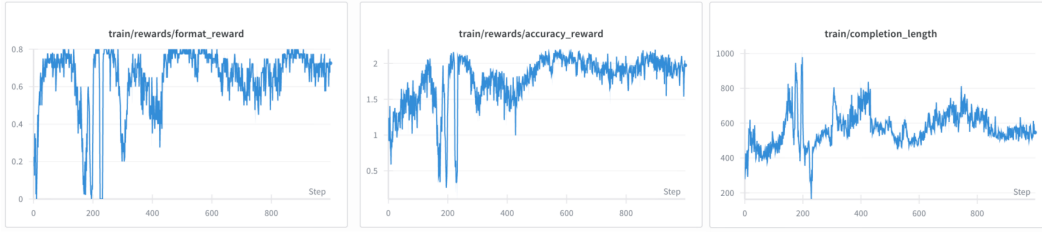


Figure 8: GRPO Training. Overall, the RL training progressed well, but there is a clear collapse around 200 steps. The trigger was the length-reward design: during exploration, longer completions earned higher scores, but once a response exceeded 600 tokens, a linear decay penalty kicked in. The policy reacted by abruptly shortening completions to 200 tokens; these outputs were often incomplete, so the format 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.

A.2 HARDER AUDIO REASONING TASKS

Following this, we further analyzed erroneous predictions of Qwen2-Audio on the MMAU benchmark. As shown in Fig.7(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

$$\text{SNR}_{\text{dB}} = 10 \log_{10} \left(\frac{P_s}{P_{n,\text{scaled}}} \right) = 10.$$

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

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

English Answer	Chinese Answer	Paraformer	Qwen-Audio	Qwen2-Audio	Qwen2-Audio+ Bad Thought	Analysis
Cuella	库艾拉	屈埃拉	Quoi elle.	Quoi elle.	屈埃拉	For very short examples, Qwen2-Audio and Qwen-Audio will do translation automatically.
Mill Creek is an unincorporated area located in Pope County, Arkansas, United States.	米尔克里克市是位于美国阿肯色州波普县的一个非建制地区。	米尔克里克市是位于美国阿肯色州波普县的一个非建制力区	米尔克里克市是位于美国阿肯色州波普县的一个非建制郡。	米尔克里克市是位于美国阿肯色州波普县的一个非建制地区。	米尔克里克市是位于美国阿肯色州波普县的一个非建制力区。	(1) Homophones are difficult to distinguish; (2) The results of the qwen2audio model are easily misled by bad answers; The term "unincorporated county" is semantically contradictory; "Bopu County" has clearly defined its administrative level as "county", so its subordinate cannot be a "county".
Caul fat is a thin membrane surrounding the internal organs of animals, used as a cooking ingredient.	网油是围绕动物内部器官的薄膜为烹饪食材	网游是围绕动物内部器官的薄膜为烹饪食材	网油是围绕动物内部器官的薄膜为烹饪食材	网油是围绕动物内部器官的薄膜为烹饪食材	网油是围绕动物内部器官的薄膜为烹饪食材 (模型输出正确, 无需修正)	Qwen2-Audio and Qwen-Audio are better at capturing contextual information compared to Paraformer.
Night scenery is really hard to capture	夜景真的超难拍	液晶真的超难拍	液晶真的超难拍	液晶真的超难拍	液晶真的超难拍	But hard cases still cannot be accurately identified. "Night scene" and "liquid crystal" have the same pronunciation, but the former is more suitable for the context of "taking photos".

Figure 9: Bad cases of Paraformer, Qwen-Audio and Qwen2-Audio on the simple ASR task.

B CASE STUDY

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>

Observed failure. Original CoT concludes bed from generic cues—"piece A vs. B," "bevel(ed) edge," "don't misassemble"—none of which uniquely identify a bed. This is a classic *reasoning overreach*: drawing a strong category conclusion from evidence that is common to many flat-pack builds. No category-specific anchor terms (e.g., *headboard*, *shelf*, *tabletop*, *backrest*) appear in the dialogue.

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: tabletop/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>

C 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 di-
arization 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

```

D 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.”