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

Anonymous Authors[1]

Abstract

We introduce **Long Listening Thoughts (LLT)**, a framework that scales open audio reasoning by eliciting *deliberative perception* and . To mitigate persistent perceptual hallucinations in Large Audio-Language Models, LLT synthesizes 1.7M reasoning traces grounded in *multi-aspect dense descriptions* that encompass semantic events, acoustic attributes, and temporal structures. By employing a novel *thought continuation* strategy, we induce System-2 cognitive behaviors, including verification, backtracking, and explicit "re-listening," to iteratively correct perceptual and reasoning errors. We fine-tune Qwen2.5-Omni on LLT, achieving state-of-the-art performance on complex benchmarks such as MMAR and MMAU. [**Jaeyeon:** Should update the final remark: Crucially, we demonstrate that this grounded supervision is a prerequisite for effective Reinforcement Learning, unlocking scaling laws for audio reasoning where standard SFT fails.]

1. Introduction

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

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

In parallel, test-time scaling has been shown to be effective for complex reasoning tasks in both language and vision domains (Guo et al., 2025; Jaech et al., 2024). However, explicit strategies for test-time scaling in audio reasoning remain underexplored, and naïvely applying long reasoning traces has been observed to be ineffective or even detrimental in this domain (). Prior work indicates that successful test-time scaling relies on structured cognitive behaviors such as verification and backtracking (Gandhi et al., 2025) that enable models to explore alternative hypotheses and refine their reasoning. These behaviors have been further shown to be effective in perceptually oriented visual reasoning (Liao et al., 2025; Acuna et al., 2025), yet have not been systematically incorporated into audio reasoning frameworks. These behaviors are also consistent with human auditory cognition, where perception is an active and

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Dataset	# of Reasoning Traces	Source of Audio Information	Refining Cognitive Behavior	Open Dataset	Open Model	Open Pipeline
COTA (Zhifei et al., 2025)	1.2M	Gemini-2.0-Flash	✗	✓	✓	✗
SARI (Wen et al., 2025)	32k	Qwen2-Audio	✗	✗	✗	✗
AF-Think (Goel et al., 2025)	470k	Gemini-2.0 Flash	✗	✓	✓	✗
AudioFlamingo-Sound-CoT	1.2M	LALM + LLM	✗	✓	✓	✗
AudSemThinker (Wijngaard et al., 2025)	900k	Multi-Models (Semantic; Visual)	✗	✓	✓	✓
AudioThinker (Wu et al., 2025)	40k	Qwen2-Audio	✗	✗	✗	✗
AudioMCQ (He et al., 2025)	571k	Audio Captions and Metadata (Contents; Events)	✗	✓	✓	✗
LongLisentingThoughts	1.7M	Multi-Auditory (Semantic, Acoustic; Temporal)	✓	✓	✓	✓

Table 1. Comparison of audio reasoning datasets with chain-of-thought (CoT) reasoning traces. Source of Audio Information indicates the type of audio-derived information used to construct reasoning traces. Refining Cognitive Behavior denotes whether the dataset explicitly incorporates cognitive behaviors such as verification or backtracking in the reasoning trace. Note that AudSemThinker leverages multiple models to capture semantic information (e.g., Qwen2, audio classifiers, and audio captioners), together with vision-based models. [**Jaeyeon:** Better way to handle AudSemThinker?]

iterative process involving attention reallocation, hypothesis revision, and self-monitoring under uncertainty () [**Jaeyeon:** I will add citations for this].

Motivated by these observations, we propose LongListeningThoughts (LLT), a new framework for constructing post-training datasets for audio reasoning that integrates deliberate multi-aspect perceptual grounding with explicit cognitive behaviors. Building upon existing audio CoT datasets, LLT introduces a scalable four-stage pipeline: (1) constructing multi-aspect dense audio descriptions, (2) augmenting challenging and audio-aware questions, (3) synthesizing plausible incorrect reasoning traces to enable reflection and correction, and (4) generating long reasoning traces with explicit cognitive behaviors.

Using LLT framework, we construct the post-training dataset, comprising approximately 1.7M reasoning traces that support supervised fine-tuning (SFT), online RL (e.g., GRPO), and offline preference optimization (e.g., DPO). The overall comparison with existing CoT datasets are given in Tabel 1. Through extensive analysis, we demonstrate that (1) multi-aspect perceptual grounding significantly improves reasoning quality, (2) leveraging harder questions enables more effective scaling of SFT data, and [**Jaeyeon:** Need to be updated: (3) explicitly encoding cognitive behaviors during SFT leads to stronger gains during RL]. Models trained with LLT achieve state-of-the-art performance on multiple audio reasoning benchmarks and generalize well across different model architectures and evaluation settings. Qualitative analysis show that it captures more detail and enhance its own reasoning through refining cognitive behaviors. We will open-source the dataset, trained models, and the full data construction pipeline to support reproducibility and future research.

2. Related Work

Large Audio Language Models. Recent advances in large audio–language models (LALMs) have demonstrated impressive capabilities in processing diverse sound domains, including speech, sound events, and music. These models are typically constructed by aligning pretrained audio encoders with large language models (LLMs) and training on large-scale audio–text datasets (). As a result, LALMs can follow natural language instructions and perform broad range of audio tasks, ranging from traditional tasks such as speech recognition and audio event classification tasks to open-ended question answering tasks that require more complex understanding and reasoning. Moreover, recent omni-modal models that process audio inputs alongside visual inputs by integrating encoders from multiple modalities with an LLM have also demonstrated strong performance in audio understanding (Xu et al., 2025a;b; **?**).

To systematically evaluate their capabilities, several recent benchmarks have been introduced to assess general audio understanding and reasoning in a question-answering format across a wide range of tasks (Sakshi et al., 2025; Ma et al., 2025b; Wang et al., 2025; **?**). However, analyses of these benchmarks reveal that perceptual errors account for the majority of failures, followed by reasoning errors. Moreover, benchmarks targeting fine-grained abilities further indicate that LALMs often miss subtle perceptual details, such as low-level acoustic attributes (Kim et al., 2025) and precise temporal structure (Bhattacharya et al., 2025; Wang et al., 2026). In this work, we address these limitations by incorporating more deliberate and detailed perceptual information when constructing reasoning traces, with the goal of improving the audio reasoning capabilities of LALMs.

Audio Reasoning Models. To further enhance the reasoning capabilities of large audio–language models (LALMs), recent works have proposed a variety of approaches targeting audio reasoning more explicitly. Mellow (**?**) introduce the ReasonAQA dataset to improve relatively small audio–language models, demonstrating strong performance on audio reasoning tasks. AudioFlamingo-3 (Goel et al., 2025) incorporates the AF-Think chain-of-thought (CoT) dataset (Wei et al., 2022) during its pre-training stage, enabling the model to adaptively reason when prompted. More targeted studies focus on improving the reasoning abilities of pretrained LALMs. AudioCoT (Ma et al., 2025a) systematically compares diverse chain-of-thought (CoT) prompting strategies on Qwen2-Audio (Chu et al., 2024). Several approaches construct CoT datasets for supervised fine-tuning of LALMs. Audio-Reasoner (Zhifei et al., 2025) builds the CoTA dataset with structured reasoning chains that explicitly decompose the reasoning process into planning, captioning, reasoning, and summarization stages. AudioFlamingo-Sound-CoT (**?**) alternates between a LALM and a text-only LLM to synthesize reasoning traces.

Other studies investigate combinations of SFT and RL. SARI (Wen et al., 2025) introduces a curriculum that integrates structured and unstructured supervision, while AudioThinker (Wu et al., 2025) explores alternative RL reward designs to encourage adaptive reasoning behavior. AudSemThinker (Wijngaard et al., 2025) leverages semantic descriptors derived from both audio and visual information to construct CoT datasets and apply RL on description-based reasoning tasks. AudioMCQ (He et al., 2025) introduces a large-scale multiple-choice dataset with approximately 570k examples and analyzes how questions with strong versus weak audio dependence affect the SFT–RL pipeline. Despite these advances, most existing CoT-based approaches rely on relatively limited auditory information and do not explicitly encode cognitive behaviors that are known to facilitate effective scaling with RL. In contrast, our work proposes a new post-training dataset construction framework that integrates richer, multi-aspect auditory information and explicitly models cognitive behaviors, supporting both SFT and RL training scenarios.

Compositional Logistics Enhanced Reasoning. While most recent multimodal models attempt to compress CoTs into dense modality predictions (e.g., rich captions), inducing systematic *deliberation* remains a fundamental challenge. We posit that the bottleneck in auditory reasoning is not merely a lack of compute, but a deficit in structural information. As argued in recent information-theoretic foundations, multi-steps and logistics-refined reasoning behaviors finding a solution within a search space, which is drastically reduced when the input is organized into compositional, hierarchical scaffolds (van de Koppel et al., 1996; Mota et al., 2013). In the auditory domain, this structural necessity is acute. Sound is inherently transient and temporally entangled; without explicit structural priors, a model must expend significant cognitive energy just to disambiguate overlapping acoustic events. In very recent mutlimodal reasoning approaches, *Long Perceptual Thoughts (Liao et al., 2025)*, attempt to resolve this by distilling long-form traces to bridge the gap between raw perception and grounded logic. However, these often focus on exhaustive description or contextualized visual information (e.g., bounding boxes) rather than logical logistics. We aim to study how to improve auditory reasoning by refined perception and refinement processes in this work. LongListeningThoughts explicitly encodes meta-cognitive scaffolds that prioritize the compositional logistics of sound. We synthesizes multi-aspect auditory descriptors, including spanning semantic hierarchies, acoustic textures, and temporal syntax, into a unified, low-entropy reasoning flow. This strategy transforms the reasoning process from a monolithic prediction task into a series of structured, cognitive transitions, laying the methodology described in the following section.

3. LongListeningThoughts

We propose LongListeningThoughts (LLT), a new framework designed to produce more perceptually grounded and cognitively structured reasoning traces for audio reasoning. Specifically, LLT produces reasoning traces that are both perceptually grounded and cognitively structured, explicitly modeling behaviors such as verification, backtracking, subgoal setting, and backward channeling (Gandhi et al., 2025). By integrating detailed multi-aspect audio descriptions, we provide the model with the necessary ”scaffolding” to resolve temporal and acoustic ambiguities that are otherwise lost in compressed representations.

A key departure from prior generative approaches (Zhifei et al., 2025; Goel et al., 2025; He et al., 2025) is our reliance on trace enrichment rather than de novo generation.

An overview of our method is shown in Figure 1.

Let (x, q, r^+, a) be an example from an existing CoT dataset for audio question answering, where x is the audio input, q is the question, r is the CoT response and a is the final answer. Note that r^+ and r^- denote correct and incorrect reasoning traces, respectively. Similar to approaches that encode cognitive behaviors through longer reasoning traces for the visual reasoning (Liao et al., 2025; Acuna et al., 2025), we generate three types of training pairs through our pipeline: (x, q, r^+, a), $(x, q, r^- \rightarrow r^+, a)$, and $(x, q, r^+ \rightarrow r^+, a)$. The transformations $r^- \rightarrow r^+$ and $r^+ \rightarrow r^+$ correspond to long reasoning traces, in which an initial reasoning trace is followed by refinement that incorporates explicit cognitive behaviors. Such transformations can be readily observed in the thinking blocks of large reasoning models. We leverage M_{instruct}, an instruction-following large language model, and M_{reason}, a large reasoning model, to construct a post-training dataset for the base LALM, M_{base}.

3.1. Stage 1: deliberating multi-aspect dense audio description

Perceptual errors account for the largest portion of failures in general audio understanding and reasoning benchmarks (Sakshi et al., 2025; Ma et al., 2025b; Wang et al., 2025), and LALMs are known to struggle with capturing fine-grained details beyond high-level audio events, such as low-level acoustic attributes (Kim et al., 2025) and precise temporal structure (Wang et al., 2026; Bhattacharya et al., 2025). This

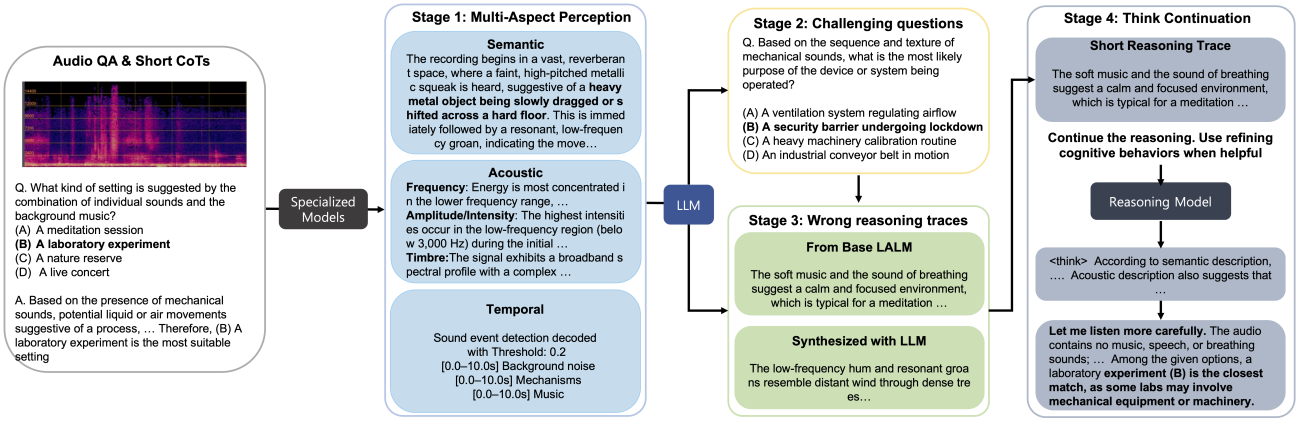


Figure 1. Overview of the proposed LongListeningThought.[**Jaeyeon:** Example might be updated with better one.][**Huck:** making the figure from png to pdf, adding emoji / small graph symblo in each stage]

suggests that improved perceptual grounding may benefit both understanding and reasoning in these models. However, existing CoT datasets typically rely on audio metadata (), or captions generated by large-audio language models themselves, such as Qwen2-Audio (), which capture only limited aspects of the audio. [**Jaeyeon:** Should add comparison with AudSemThinker: Only exception is AudSemthinker, where tried to use audio and visual analysis together, and multiple models. But qwen2 + audio caption, audio event tagging, which convey similar aspect of what events are happening.] Therefore, we propose to build a multi-aspect dense description for each audio input, providing a more holistic view of the auditory scene and richer cues for reasoning trace generation.

We divide audio information into three complementary aspects: (1) semantic information describing high-level audio events and their interpretations, (2) acoustic information capturing attributes below the event level, such as duration, pitch, timbre, and amplitude, and (3) temporal information providing precise timestamps and temporal structure. For semantic information, we use Qwen3-Omni-Captioner (Yang et al., 2025a), a recent specialized model to generate dense descriptions of audio events. For acoustic information, we directly leverage rich acoustic cues from spectrograms by prompting a vision-language model (VLM) to describe attributes such as frequency distribution, duration, spectral patterns, and reverberation. We find that providing explicit grid and axis information that the VLM can numerically ground, together with detailed interpretation guidelines (e.g., which acoustic attributes to attend to and illustrative examples), substantially improves the quality of spectrogram-based descriptions. Since large audio–language models (LALMs) typically lack a detailed understanding of low-level acoustic attributes, this spectrogram-driven approach yields more accurate and informative acoustic descriptions than relying solely on audio-language models (see Figure 1 for examples). We use Qwen3-VL-32B (Team, 2025) as the VLM for spectrogram interpretation.

For temporal information, we employ specialized models to obtain accurate timestamps and temporal structure. Specifically, Whisper-large-v3 (Radford et al., 2023) is used to generate timestamps for speech segments, while pyannote (Bredin, 2023) provides speaker diarization timestamps. In addition, we use a pretrained sound event detection model (Schmid et al., 2025) based on BEATs (Chen et al., 2023) to extract event boundaries. We apply multiple detection thresholds (0.2 and 0.5) for sound event detection in order to capture both prominent and subtle events. Finally, for audio clips detected to contain music, we additionally apply allin1 () music structure analysis and chord recognition () to obtain temporally aligned musical annotations. By concatenating all information, we form the dense description D. An example of the generated dense description is shown in Figure **??**.

3.2. Stage 2: Refining challenging questions

With a more holistic view of the auditory scene, we can generate more challenging questions grounded in richer perceptual evidence. Given an existing datapoint (x, q, r, a), we prompt M_{instruct} to generate a new datapoint $(x, q_{\text{hard}}, r_{\text{hard}}, a_{\text{hard}})$. Specifically, the model is instructed to retain the original question format and required skill set, while increasing difficulty by making the question more audio-aware or by requiring more complex reasoning. If this is not feasible, the model is allowed to generate a new question grounded in the dense audio description.

Additionally, we verify the generated questions using M_{LLM} to ensure that (1) the question can be solved by listening to the audio, (2) the final answer is correct, and (3) the associated reasoning trace is logically valid.

3.3. Stage 3: Generating wrong reasoning traces

To enhance reflective and corrective reasoning, we first generate incorrect reasoning traces r^-. As an initial source of errors, we include wrongly generated reasoning traces produced by the base model M_{base}, which allows us to directly capture and correct failure patterns that M_{base} exhibits in practice. However, relying solely on such naturally occurring errors has two limitations: (1) we find out that incorrect reasoning traces are not always generated from M_{base} even when the model is explicitly prompted to reason (Wu et al., 2025), and (2) the resulting errors lack sufficient diversity to reflect the range of misperceptions that can arise in real-world scenarios. When we randomly sampled 80k QAs from origianl CoT dataset we used (Goel et al., 2025) and ask Qwen2.5-omni-7B (Xu et al., 2025a) to generate reasoning trace, only 11k of wrong reasoning trace could be obtained.

To simulate more diverse and realistic perceptual failures, we additionally generate synthetic wrong reasoning traces by prompting a language model with (q, r^+, a^+, D). Specifically, we instruct the model to retain the length and structural format of the correct reasoning trace r^+, while generating a plausible but incorrect reasoning trace r^- that arises from misperceiving or misinterpreting fine-grained details in the dense audio description. This process encourages the generation of errors that are subtle, realistic, and grounded in perceptual ambiguity.

We generate r^- for both original questions and the hardened questions produced in Stage2. We combine both naturally occurring incorrect traces and synthetically generated incorrect traces. All resulting incorrect reasoning traces are subsequently used in Stage4 to generate long reasoning traces with explicit cognitive behaviors.

3.4. Stage 4: Generating long reasoning traces through thought continuation

Previous work in the vision domain adopts a thought expansion approach to extend reasoning traces, typically by directly distilling long reasoning traces from a large reasoning model M_{reason} (Liao et al., 2025; Acuna et al., 2025). Specifically, this approach combines a short reasoning trace r with cognitive trigger tokens such as "Wait," "Hmm," or "Alternatively," and prompts M_{reason} to generate a completed reasoning trace within the `<think></think>` block, conditioned on the question and the visual description.

However, directly applying this strategy to audio reasoning is ineffective due to the length and multi-aspect nature of dense audio descriptions. M_{reason} sequentially referred to the different types of information and often struggles to determine which parts of the description to attend to, leading to excessively long, unfocused, or repetitive reasoning traces. Such behavior is undesirable for . Therefore, instead of directly distilling full reasoning traces from M_{reason}, we adopt a thought-continuation approach.

Specifically, given $(q, r \oplus c, D)$, where $r \oplus c$ denotes a short reasoning trace followed by one of the randomly selected cognitive trigger and D is the dense audio description, we prompt the M_{reason} to generate a continuation of the $r \oplus c$ as the output of thinking. In other words, this continuation is generated as the part of `<answer></answer>` following the `<think></think>` block, therefore M_{reason} itself refine its own reasoning to solve the question and make it natural form to follow original format. To further ensure the, we explicitly prompt the model to encode cognitive behaviors such as verification and backtracking within this continuation. In addition, we incorporate audio-aware cues (e.g., `Let me listen again`, `Let me listen carefully`) together with general cognitive cues (e.g., `Wait`, `Hmm`, `Alternatively`, `Let me check again`) to encourage stronger grounding in the audio signal.

This approach enables the encoding of cognitive behaviors and longer reasoning traces inspired by reasoning models, while producing outputs that are relatively concise, refined, and better suited for SFT compared to directly distilling raw thinking tokens. [**Jaeyeon:** Maybe add brief comparison on lengths between reasoning traces from 1. origianl 2. thought expansion 3. thought continuation]. An example comparing direct thought expansion with our thought-continuation approach is shown in Figure **??**. We apply thought continuation to both z^+ traces from the original dataset and Stage2, as well as z^- traces from Stage3, resulting in long reasoning traces of the forms $(r^+ \to r^+)$ and $(r^- \to r^+)$.

In addition, we apply verification during long reasoning trace generation. Specifically, we prompt M_{instruct} with the generated continuation, and (q, a, D), and verify the following criteria: (1) whether the final answer is correct, (2) whether the reasoning is logically valid, and (3) whether the output format is valid (e.g., no references to the description text and proper use of the `<answer>` format). For criteria (1) and (2), samples that fail verification are discarded. For criterion (3), minor formatting errors are corrected, while samples with severe format violations are discarded.

3.5. Implementation Details.

We use AF-Think as the initial short CoT dataset, as it covers diverse audio domains and scenarios. For reproducibility, we exclude licensed data such as Fisher () and Music4all (), and limit audio duration to 30 seconds for compatibility with general LALM architectures, resulting in 370k audio clips and corresponding original question–answer pairs per audio. Through Stage 2, we augment 320k challenging questions, resulting in a total of 690k QA pairs. Following

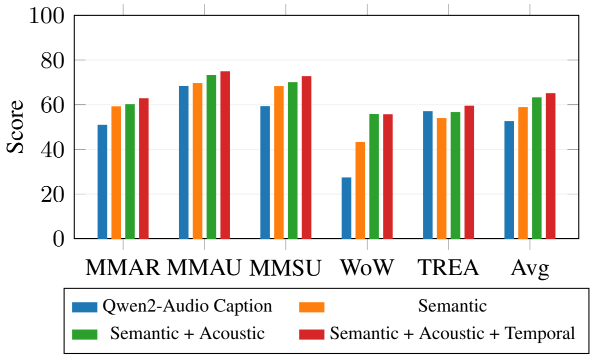


Figure 2. Accuracy of an LLM across different benchmarks using different types of audio descriptions. Semantic, Acoustic, and Temporal correspond to the respective components of the multi-aspect audio descriptions constructed in Stage 1 of LLT.

AF-Think, which includes both multiple-choice and free-form questions, we retain both formats in our dataset. We retain 4k audios and corresponding qas for valdiation, and 4k for further GRPO training, resulting 650k QA for SFT training.

Based on these QA pairs, the final LLT has 650k (x, q, r^+, a) pairs, $(x, q, r^- \to r^+, a)$ 460k pairs, and 550k $(x, q, r^+ \to r^+, a)$ pairs, resulting in a total of 1.66M training instances for supervised fine-tuning (SFT). Throughout the pipeline, we use Qwen3-Next-80b-a3b-fp8-instruct as M_{instruct} and Qwen3-Next-80b-a3b-fp8-think M_{reason}. All components of the pipeline are based on open-source models, including the submodules used in Stage1, ensuring full reproducibility. The prompts used at each stage are provided in Appendix **??**.

We additionally utilize LLT for reinforcement learning (RL) training, enabling a two-stage training procedure consisting of SFT followed by RL. For GRPO, we sample 4k QA pairs in multiple-choice format. We restrict GRPO training to the MCQ setting to enable straightforward rule-based reward design, following common practice (**?**). We further construct ?? preference pairs for DPO training. [**Jaeyeon:** Need to add details for DPO pairs]

4. Analysis on Audio Reasoning

Through controlled analyses using AF-Think and LLT, we identify several important findings for building effective CoT datasets for audio reasoning. Specifically, we observe that (1) richer perceptual information enables better reasoning traces, (2) increased question complexity facilitates more effective scaling of SFT data, and (3) explicitly encoding reasoning behaviors during SFT is critical for downstream RL performance.

Does richer perceptual information benefit audio reasoning? To evaluate how additional perceptual information affects audio reasoning, we use a strong audio–language

Model	Accuracy w/ Audio	Accuracy w/o Audio	Gap
Original	72.5	45.5	27
LLT Stage 2	63.1	55.6	7.5

Table 2. Comparison of model accuracy with and without audio input.

reasoning setup consisting of detailed audio descriptions and Qwen3-32B (Yang et al., 2025a) in thinking mode. We evaluate the model under different description settings by prompting it to answer audio question–answer pairs on both general audio understanding and reasoning benchmarks (Sakshi et al., 2025; Ma et al., 2025b; Wang et al., 2025), as well as specialized benchmarks targeting acoustic properties (Kim et al., 2025) and temporal reasoning (Bhattacharya et al., 2025).

Specifically, we compare the widely used Qwen2-audio based captions, the more detailed Qwen3-Omni-Captioner caption corresponding to the semantic information from Stage1 of our pipeline, and progressively richer descriptions obtained by additionally incorporating acoustic and temporal information from Stage1. The results are shown in Figure 2 We find that using more specialized models to generate richer descriptions consistently improves performance. Moreover, adding acoustic and temporal information not only improves performance on their corresponding specialized benchmarks, but also leads to gains on general multi-task benchmarks. These results indicate that the proposed multi-aspect descriptions provide more effective perceptual grounding and supply sufficient cues to generate high-quality reasoning traces for text-based LLMs on audio-related questions.

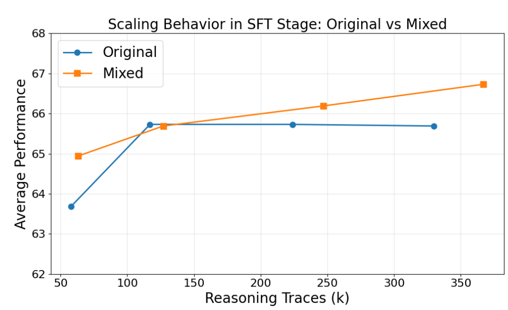


Figure 3. Scaling behavior.[**Jaeyeon:** Need to add captions]

How does question quality affect SFT scaling? To study the effect of question quality, we randomly select 10k audio clips from the original AF-Think dataset and compare the accuracy of Qwen2.5-Omni-7B on the original question–answer pairs and the corresponding hardened question pairs generated by LLT. The model achieves an accuracy of 72.5% on the original questions, while accuracy drops to 63.0% on the hardened questions. This performance gap indicates that the LLT pipeline successfully generates more

Method	MMAU test-mini	MMAR	MMSU
SFT with 50k QA Pairs			
AF-Think	74.1	58.8	63.3
+ $r^+ \to r^+$	**75.7**	**61.1**	**64.0**
+ $r^- \to r^+$	75.0	58.9	63.3
SFT with Full Reasoning Traces			
AF-Think	73.9	59.9	63.7
+ $r^+ \to r^+$	**76.3**	**62.2**	**65.0**
+ $r^- \to r^+$	76.2	61.9	64.9
GRPO After Full SFT			
AF-Think			
+ $r^+ \to r^+$			
+ $r^- \to r^+$			

Table 3. Ablation study on training components and reasoning trace construction.

complex and challenging questions that require deeper audio reasoning. [**Jaeyeon:** Need to add explnation for small gap with w/o audio setup: requires more semantic reasoning]

We further analyze how question complexity influences SFT scaling. Using the same audio inputs, we construct two SFT datasets: one consisting of half hardened questions and half original questions, and another consisting entirely of original AF-Think questions. We used LLT reasoning pairs for both cases. When scaling the audio pairs size from 25k to 150k, which corresponds to the reasoning traces about 50k to more than 300k, we observe consistent performance gains when training on the mixed dataset compared to train only on the original quesitons, as shown in Figure 3. This suggests that question complexity plays an important role in effectively scaling SFT data, even when the underlying audio content remains unchanged. [**Jaeyeon:** Explanation for the performance axis: Average MMSU / MMAU-Test min / MMAR.]

Does explicitly encoding cognitive behaviors benefit audio reasoning?[**Jaeyeon:** I am still running the RL experiments, and content of this section might be changed.]

In previous works, SFT -¿ RL benefits is observed. However, the . 1. wrong CoT -¿ degrades SFT performance compared to using only o -¿ o. This means that refining helps. However, due to limited reflective action, yield much less cognitive behaviors. This gap between smaller when trained on larger scale, model learns to follow wrong CoT or not.

After RL, changes: Cogntivie behavior helps performance boost.

5. Experiments

5.1. Setup

Training Details. Following prior work (Wijngaard et al., 2025; Wu et al., 2025; He et al., 2025; **?**), we train Qwen2.5-Omni-7B on the proposed LLT dataset. We perform full fine-tuning of all model components, including the LLM, the audio encoder, and the projection layers. For SFT, we use a learning rate of 6×10^{-6} and train the model for 5 epochs. All SFT experiments are implemented using LLaMAFactory (Zheng et al., 2024). RL is performed starting from the SFT-trained checkpoint. For GRPO, we use a learning rate of 1×10^{-6}, generate 8 samples per prompt, and set the batch size to 16. We use a KL coefficient β of 0.005 and a sampling temperature of 1.5. RL training is implemented using Swift (), with 4k questions and 2,000 optimization steps. We adopt a rule-based reward scheme following common practice. We adopt an accuracy reward of 1 is assigned if the multiple-choice answer is correct and 0 otherwise, and an additional format reward is given for outputs that correctly follow the `<think></think>` and `<answer></answer>` tag structure. The accuracy reward is weighted by 1.0, and the format reward is weighted by 0.05.

Baselines. To isolate the effect of the LLT dataset, we compare against a Qwen2.5-Omni-7B model trained on the original AF-Think dataset using the same set of audio inputs. In addition, we compare our approach with recent large audio–language models and audio reasoning models that demonstrate strong performance on benchmark evaluations. We note that many prior audio reasoning models are also built upon the Qwen2.5-Omni-7B backbone, enabling a fair comparison.

Evaluation Details. We evaluate all models on the widely used MMAU (Sakshi et al., 2025), MMAR (Ma et al., 2025b), and MMSU (Wang et al., 2025) benchmarks, which assess general audio understanding and reasoning capabilities across diverse tasks. MMAU evaluates information extraction and reasoning across speech, sound events, and music. We use the most recent MMAU-test-v0515 with refined question and answers and report results from the official leaderboard whenever available. MMAR focuses on complex, multi-step reasoning tasks spanning speech, sound, music, and their combinations. MMSU places greater emphasis on speech, evaluating both perceptual and reasoning abilities across paralinguistic and linguistic dimensions. For all benchmarks, we follow the official evaluation protocols: string matching is used for MMAU and MMAR, and letter matching is used for MMSU, with answers extracted from the `<answer></answer>` tags.

Method	MMAU test-mini	MMAU	MMAR	MMSU
Large Audio Language Models				
Audio Flamingo 3	73.3	72.4	58.5	61.4
+ Think	74.3	73.2	60.1	62.3
Kimi-Audio	68.2	64.4	57.6	59.3
Audio Reasoning Models				
Audio-Reasoner	67.7	63.8	36.8	49.2
R1-AQA	68.9	68.5	50.8	61.6
SARI	67.0	–	–	66.0
Omni-R1	77.0	75.0	63.4	–
Audio-Thinker	78.0	75.4	65.3	–
AudioMCQ	78.2	75.6	65.3	69.3
AudioMCQ-2	75.8	74.6	67.1	70.7
Proprietary Models				
GPT4o-Audio	62.5	60.8	63.5	56.4
Gemini-2.0-Flash	70.5	67.0	65.6	51.0
Our Methods				
Qwen2.5-Omni	71.5	71.0	56.7	60.6
+ AF-Think SFT	73.9		59.9	63.7
+ LLT SFT	76.2		61.9	64.9
+ LLT SFT + DPO	76.9		62.6	66.1
+ LLT SFT + GRPO				

Table 4. Performance comparison across different models.[**Jaeyeon:** Number of our variants are from 120k samples]

Method	MMAU test-mini	MMAR	MMSU
Training Components			
LLM only	74.2	**60.0**	63.0
+ Projector	74.9	59.4	62.4
+ Audio Encoder	**75.0**	58.9	**63.3**
Stage 4: Reasoning Trace			
Thought expansion			
Thought continuation	75.0	58.9	63.3

Table 5. Ablation study on training components and reasoning trace construction.

5.2. Results

Comparison to baselines. Outperform original AF-Think. Performance further scale up with RL. Outperform all other models. Table 4

Ablation on the LLT pipeline. We conduct ablation using 50k audios and corresponding 117k LLT reasoning pairs for the ablation. We conduct only the SFT. Tabel 5

For freezing different components across the model. LLM only best perforamnce on MMAR, which requires heaviest reasoning among this benchmarks. However, training encoder and projection together further imporves MMAU and MMSU which also has perception related subtasks,

Stage 3: [**Jaeyeon:** Having only correct traces perform better on SFT... How should I handle this?: I will have analysis on this based on the cognitive behavior]

Stage 4:

Model	WoW-Bench	TREA	Omni-Bench	MMLU
Qwen2.5-Omni				
+ LLT SFT				
+ LLT SFT + GRPO				

Table 6. Perforamnce across different benchmarks

Method	MMAU test-mini	MMAR	MMSU
Opus-LM-3b			
+ LLT SFT			
Qwen2-Audio-Instruct	59.6	30.0	53.3
+ LLT SFT	**66.4**	**52.2**	**56.5**

Table 7. Perfomance across different models.

Generalization

general perceptual ability: wow / trea

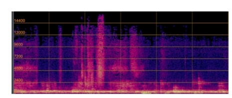
Reasoning ability generlize across domain: omni-bench, mmluT

Table 6

Can be applied to different audio model: opus-lm. qwen2-audio. Trained with 120k trace from LLT, improve performance.

Table 7

Qualitative Examples



Q. Is the audio the original sound of a football match, if not then what activity is it from?
A. No, it's a gym workout B. It's a football match
C. No, it's a track and field event **D. No, it's training audio**

(Qwen2.5-Omni-7B + AF-Think)
<think> The audio contains sounds of a football match, including a whistle and a goal being scored. Therefore, the answer is: (B) It's a football match. </think>

<answer> **(B) It's a football match** </answer>

(Qwen2.5-Omni-7B + LLT)
<think> The audio contains sounds of a dog barking and a man speaking, which could resemble a referee's whistle or crowd noise. The mention of 'good' and 'nice' might be interpreted as cheers or encouragement. These elements suggest a live sports event, making it likely a football match. **Wait. The audio features a male voice giving commands like "Bloom" and "Shut," paired with a dog's vocalizations and a metallic click, which aligns more closely with a structured training session** than a live sports event. The absence of crowd noise, the focus on precise physical movements (e.g., "like you're sitting on a toilet"), and the use of motivational phrases ("Good," "Nice") suggest a coaching context rather than a competitive match. The metallic click likely corresponds to a training tool or equipment, not a referee's whistle. These details point to a training scenario. </think>

<answer> **(D) No, it's training audio** </answer>"

Figure 4. Qualitative Example.

count of cognitive behaviors / response length

real reasoning comparison

more qualitative examples at appendix

Long-Horizon Thinking in Audio Understanding and Its Error Types

Cognition Categories

What wins

What loss

6. Conclusion

Concluding remarks.

Impact Statement

This work introduces LongListeningThoughts (LLT), a framework that advances the reasoning capabilities of Large Audio-Language Models (LALMs) by integrating deliberative perception with cognitive refinement. While the primary goal of this research is to mitigate perceptual hallucinations and improve fine-grained auditory understanding, its deployment carries significant societal implications.

Applications. LLT introduces a scalable pipeline to synthesize 1.7M reasoning traces grounded in multi-aspect audio descriptions. This framework enables several key properties in downstream applications: (1) Enhanced Reliability via System-2 cognitive behaviors like verification and backtracking; (2) High-Fidelity Perception that grounds reasoning in low-level acoustic attributes and temporal structures; and (3) Scalability through a trace-enrichment strategy that is backward-compatible with legacy datasets.

Implications on Societal Outcomes and Risks The transition from simple audio labeling to deliberative reasoning has profound implications for **accessibility technology**. For the visually impaired, LLT-powered systems can provide nuanced context—distinguishing, for example, a structured training session from a live sports event based on subtle acoustic cues. Improved temporal reasoning is also critical for **safety-oriented applications**, such as automated industrial diagnostics.

We identify potential **surveillance and privacy risks**. LLT's ability to perform speaker diarization and extract semantic hierarchies could be misused to extract sensitive personal metadata from private audio recordings without authorization. Furthermore, there is a risk of **hallucinated "logic"** in high-stakes environments, where the "thought-continuation" strategy might lead models to provide logically consistent but factually incorrect justifications for their decisions.

[**Huck:** Adding discussion on descriptive bias (Hirota et al., 2024) and de-bias is needed for commercial applications.]

References

Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Acuna, D., Yang, C.-H. H., Deng, Y., Jung, J., Lu, X., Ammanabrolu, P., Kim, H., Liao, Y.-H., and Choi, Y. Long grounded thoughts: Distilling compositional visual reasoning chains at scale. *arXiv preprint arXiv:2511.05705*, 2025.

Bhattacharya, D., Kulkarni, A., and Ganapathy, S. Benchmarking and confidence evaluation of lalms for temporal reasoning. In *Interspeech*, 2025.

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

Chen, C., Hu, Y., Wang, S., Wang, H., Chen, Z., Zhang, C., Yang, C.-H. H., and Chng, E. Audio large language models can be descriptive speech quality evaluators. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., and Wei, F. Beats: Audio pre-training with acoustic tokenizers. In *ICML*, 2023.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Deshmukh, S., Elizalde, B., Singh, R., and Wang, H. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108, 2023.

Gandhi, K., Chakravarthy, A., Singh, A., Lile, N., and Goodman, N. D. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

Ghosh, S., Kumar, S., Seth, A., Evuru, C. K. R., Tyagi, U., Sakshi, S., Nieto, O., Duraiswami, R., and Manocha, D. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *EMNLP*, 2024.

Ghosh, S., Goel, A., Kim, J., Kumar, S., Kong, Z., Lee, S.-g., Yang, C.-H. H., Duraiswami, R., Manocha, D., Valle, R., et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Goel, A., Ghosh, S., Kim, J., Kumar, S., Kong, Z., Lee, S.-g., Yang, C.-H. H., Duraiswami, R., Manocha, D., Valle, R., et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

Gong, Y., Luo, H., Liu, A. H., Karlinsky, L., and Glass, J. Listen, think, and understand. In *International Conference on Learning Representations*, 2024.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

He, H., Du, X., Sun, R., Dai, Z., Xiao, Y., Yang, M., Zhou, J., Li, X., Liu, Z., Liang, Z., et al. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. *arXiv preprint arXiv:2509.21060*, 2025.

Hirota, Y., Hachiuma, R., Yang, C.-H., and Nakashima, Y. From descriptive richness to bias: Unveiling the dark side of generative image caption enrichment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17807–17816, 2024.

Hu, Y., CHEN, C., Yang, C.-H. H., Li, R., Zhang, C., Chen, P.-Y., and Chng, E. Large language models are efficient learners of noise-robust speech recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Kim, J., Yun, H., Woo, S. H., Yang, C.-H. H., and Kim, G. Wow-bench: Evaluating fine-grained acoustic perception in audio-language models via marine mammal vocalizations. *arXiv preprint arXiv:2508.20976*, 2025.

Li, G., Liu, J., Dinkel, H., Niu, Y., Zhang, J., and Luan, J. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025.

Liao, Y.-H., Elflein, S., He, L., Leal-Taixé, L., Choi, Y., Fidler, S., and Acuna, D. Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception. *arXiv preprint arXiv:2504.15362*, 2025.

Lu, K.-H., Chen, Z., Fu, S.-W., Yang, C.-H. H., Huang, S.-F., Yang, C.-K., Yu, C.-E., Chen, C.-W., Chen, W.-C., Huang, C.-y., et al. Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *arXiv preprint arXiv:2507.02768*, 2025.

Ma, Z., Chen, Z., Wang, Y., Chng, E. S., and Chen, X. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*, 2025a.

Ma, Z., Ma, Y., Zhu, Y., Yang, C., Chao, Y.-W., Xu, R., Chen, W., Chen, Y., Chen, Z., Cong, J., et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025b.

Mota, F., Aaronson, S., Antunes, L., and Souto, A. Sophistication as randomness deficiency. In *International Workshop on Descriptional Complexity of Formal Systems*, pp. 172–181. Springer, 2013.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Radhakrishnan, S., Yang, C.-H. H., Khan, S. A., Kumar, R., Kiani, N. A., Gomez-Cabrero, D., and Tegner, J. N. Whispering llama: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10007–10016, 2023.

Sakshi, S., Tyagi, U., Kumar, S., Seth, A., Selvakumar, R., Nieto, O., Duraiswami, R., Ghosh, S., and Manocha, D. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

Schmid, F., Morocutti, T., Foscarin, F., Schlüter, J., Primus, P., and Widmer, G. Effective pre-training of audio transformers for sound event detection. In *ICASSP*, 2025.

Team, Q. Qwen3-vl: Sharper vision, deeper thought, broader action. https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef, 2025. Accessed: 2025-11-24.

van de Koppel, J., Huisman, J., van der Wal, R., and Olff, H. Patterns of herbivory along a prouductivity gradient: An empirical and theoretical investigation. *Ecology*, 77(3): 736–745, 1996.

Wang, D., Wu, J., Li, J., Yang, D., Chen, X., Zhang, T., and Meng, H. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*, 2025.

Wang, H., Li, Y., Ma, S., Liu, H., and Wang, X. Listening between the frames: Bridging temporal gaps in large audio-language models. In *AAAI*, 2026.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Wen, C., Guo, T., Zhao, S., Zou, W., and Li, X. Sari: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*, 2025.

Wijngaard, G., Formisano, E., Esposito, M., and Dumontier, M. Audsemthinker: Enhancing audio-language models through reasoning over semantics of sound. In *NeurIPS*, 2025.

Wu, S., Li, C., Wang, W., Zhang, H., Wang, H., Yu, M., and Yu, D. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*, 2025.

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yang, C.-H. H., Ghosh, S., Wang, Q., Kim, J., Hong, H., Kumar, S., Zhong, G., Kong, Z., Sakshi, S., Lokegaonkar, V., et al. Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge. *arXiv preprint arXiv:2505.07365*, 2025b.

Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., and Ma, Y. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *ACL Demo*, 2024.

Zhifei, X., Lin, M., Liu, Z., Wu, P., Yan, S., and Miao, C. Audio-reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23840–23862, 2025.

A. Examples from LLT pipeline

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

A.1. Acosutic Description Example

A.2. Stage 1 Description Example

A.3. Though Expansiion vs. Though Continuation

B. Prompts for LLT

B.1. Stage 1

¡THINK¿
¡PLANNING¿
The user asks me to identify the number of speakers in the audio clip. To do this, I will listen carefully and try to distinguish unique voices. The conversation is between two hikers, I need to confirm this by identifying their individual lines and vocal characteristics. Finally, I will provide the answer based on the number of identified speakers.
¡/PLANNING¿
¡CAPTION¿
The audio presents a conversation between two individuals hiking.
Speaker 1: This person starts by commenting on the intensity of the climb and expresses doubt about continuing without a break. Later, after a reminder about the ranger's warning, this speaker agrees to turn back. The vocal tone suggests exertion and a desire to reach the summit, eventually giving way to agreement. Phrases spoken: "Wow, this climb is intense. I am not sure I can go much further without a serious break.", "Turn back. But we're so close to the summit. I really want to make it, you know.", "Unpredictable mountain weather, okay. Okay, you are right."
Speaker 2: This person responds to the first speaker's concerns, suggesting they turn back due to the darkening sky and diminishing daylight. They emphasize safety and remind the other speaker about the ranger's warning. The vocal tone conveys concern and prioritizes safety. Phrases spoken: "I know right. I am starting to think we should turn back. That sky looks awfully dark, and we're losing daylight quickly.", "I hear you, but safety first. Remember what the ranger said about the unpredictable mountain weather.", "So, turn back."
Based on the clear distinction in vocal tones and the back-and-forth nature of the dialogue, I can confirm that there are two distinct speakers in this audio clip.
¡/CAPTION¿
¡REASONING¿
Initial Assessment: The audio starts with two distinct voices in a hiking - related dialogue. The voices are clearly distinguishable from the get - go, with one having a deeper timbre and the other a higher - pitched one.
Dialogue Analysis: A close look at the content shows a clear difference in the speakers' stances. One is eager to press on to the summit, saying, "We're almost there. Let's keep going." The other is more cautious, citing the darkening sky and fading daylight, stating, "The weather looks bad. We should turn back."
Vocal Tone Differentiation: Each speaker's vocal tone stays consistent. The eager speaker has an energetic and determined tone, with a rising pitch when advocating for continuing. The cautious speaker has a measured, worried tone, with a lower pitch when warning of danger.
Confirming Speaker Count: A review of the audio shows no extra voices or interruptions. The conversation moves smoothly between the two speakers, indicating only two are present.
Final Answer: Based on this analysis, there are two individuals in the conversation.
Thus, the answer is B.
¡/REASONING¿
¡SUMMARY¿
Based on the dialogue and vocal tone analysis, there are two distinct speakers in the audio. The answer is B.
¡/SUMMARY¿

¡/THINK¿
¡RESPONSE¿
Based on the dialogue and distinct vocal tones present in the audio, I have identified two individuals engaged in the conversation. Therefore, the answer is B.
¡/RESPONSE¿

B.2. Stage 2

B.3. Stage 3

B.4. Stage 4

C. More qualitative examples

D. Detailed results for each benchmark