LongListeningThoughs: Eliciting Open Auditory Reasoning with Deliberative Perception and Cognitive Refinement

Anonymous Authors[1]

Abstract

We introduce **LongListeningThoughts (LLT)**, a framework that scales open audio reasoning by eliciting *deliberative perception*. To mitigate persistent perceptual hallucinations in Large Audio-Language Models, LLT synthesizes 1.4M 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 errors. We fine-tune Qwen2.5-Omni on LLT, achieving state-of-the-art performance on complex benchmarks such as MMAR and MMAU. 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

[**Jaeyeon:** TODO: Placeholder from final report. Should rewrite entirely] Recent progress in large audio-language models (LALMs) has significantly advanced audio processing across a variety of tasks, yet their reasoning abilities remain relatively limited. In particular, perceptual errors constitute the majority of failures in audio reasoning benchmarks (Sakshi et al., 2025; **?**), and there has been little exploration of test-time scaling for audio reasoning despite its demonstrated success in NLP and vision domains.

To address these limitations, we propose LongListeningThoughts (LLT), a new dataset-generation pipeline that constructs long, cognitively structured Chain-of-Thought (CoT) (Wei et al., 2022) traces grounded in detailed audio descriptions. Our approach first enriches each example with multi-level perceptual information using Qwen3-Omni-Captioner (Yang et al., 2025) and spectrogram-based prompting using a vision-language model (VLM) (Team, 2025). We then extend short CoTs from AF-Think (Goel et al., 2025) into long-form reasoning traces using a large language model(LLM) with strong reasoning capabilities (**?**), incorporating behaviors such as verification and backtracking. These traces provide both richer supervision during training and stronger perceptual grounding.

2. Related Work

Large Audio-Language Models.

Audio Reasoning Models.

Test-Time Scaling.

3. LongListeningThoughts

We propose LongListeningThoughts (LLT), a new framework designed to produce more perceptually grounded and cognitively structured reasoning traces for audio reasoning. These traces explicitly model cognitive behaviors such as verification, backtracking, subgoal setting, and backward channeling (Gandhi et al., 2025), while enhancing perceptual grounding through the integration of multi-aspect, detailed audio descriptions. Unlike existing methods that construct audio question–answer pairs and CoT datasets from scratch (), LLT operates by extending and enriching existing CoT traces, making it scalable and compatible with datasets from prior approaches (Zhifei et al., 2025; Goel et al., 2025; He et al., 2025). 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 $(x, q, r^+ \rightarrow r^+, a)$.

[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	Explicit Cognitive Behavior	Open Dataset	Open Model	Open Pipeline
COTA (Zhifei et al., 2025)	1.2M					
SARI		Qwen2-Audio-Instruct				
AF-Think (Goel et al., 2025)	470k	Gemini-2.0 Flash	✗	✓	✓	✗
AudioFlamingo-Sound-CoT	1.2M		✗			
AudSemThinker						
AudioThinker (Wu et al., 2025)						
AudioMCQ (He et al., 2025)				✓	✓	✗
LongLisentingThoughts	1.4M	Multi-Aspect	✓	✓	✓	✓

Table 1. Comparison of audio reasoning datasets. [**Jeff:** Can you fill in the table and search for additional entries if exists?]

The transformations $r^- \to r^+$ and $r^+ \to r^+$ 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., 2025; 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 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., 2025), 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,

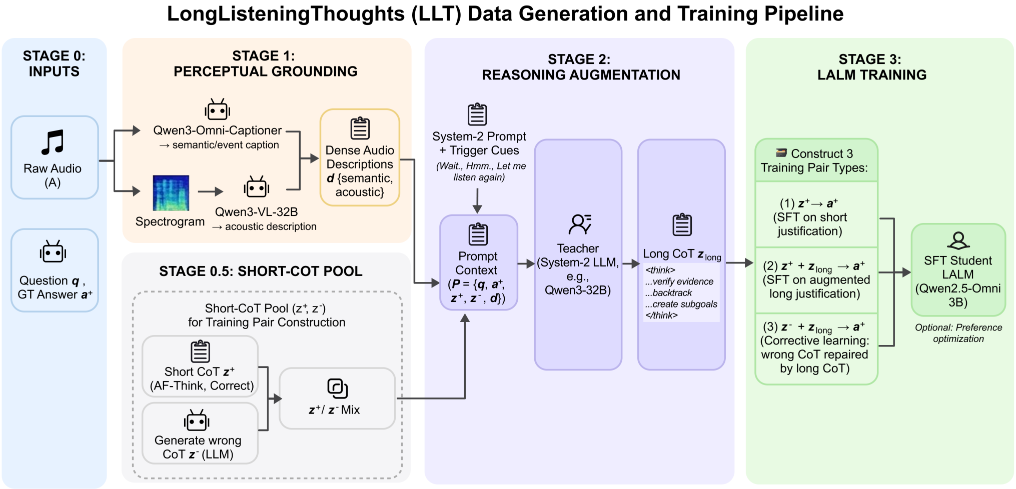
LongListeningThoughts (LLT) Data Generation and Training Pipeline
STAGE 0: INPUTS
Raw Audio (A)
Question q, GT Answer a^+
STAGE 1: PERCEPTUAL GROUNDING
Qwen3-Omni-Captioner → semantic/event caption
Spectrogram
Qwen3-VL-32B → acoustic description
Dense Audio Descriptions d {semantic, acoustic}
STAGE 0.5: SHORT-COT POOL
Short-CoT Pool (z^+, z^-) for Training Pair Construction
Short CoT z^+ (AF-Think, Correct)
Generate wrong CoT z^- (LLM)
z^+/ z^- Mix
STAGE 2: REASONING AUGMENTATION
System-2 Prompt + Trigger Cues
(Wait., Hmm., Let me listen again)
Prompt Context (P = {q, a^+, z^+, z^-, d})
Teacher (System-2 LLM, e.g., Qwen3-32B)
Long CoT z_{long}
<think>
...verify evidence
...backtrack
...create subgoals
</think>
STAGE 3: LALM TRAINING
Construct 3 Training Pair Types:
(1) $z^+ \rightarrow a^+$ (SFT on short justification)
(2) $z^+ + z_{long} \rightarrow a^+$ (SFT on augmented long justification)
(3) $z^- + z_{long} \rightarrow a^+$ (Corrective learning: wrong CoT repaired by long CoT)
SFT Student LALM (Qwen2.5-Omni 3B)
Optional: Preference optimization

Figure 1. [**Jaeyeon:** Placeholder. Should be replaced with the final figure.] Overview of the proposed LongListeningThought data generation and training pipeline.

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 solely 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., 2025) 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 $(r^+ \rightarrow r^+)$ and $(r^- \rightarrow r^+)$.

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.

[**Jaeyeon:** Numbers to be filled based on final status.] 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 ?? audio clips and question–answer pairs. Through Stage 2, we augment ?? challenging questions, resulting in a total of ?? QA pairs. Following AF-Think, which includes both multiple-choice and free-form questions, we retain both formats in our dataset.

Based on these QA pairs, the final LLT has ??? (x, q, r^+, a) pairs, $(x, q, r^- \rightarrow r^+, a)$?? pairs, and $(x, q, r^+ \rightarrow r^+, a)$ pairs, resulting in a total of ?? 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 randomly sample ?? 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. While additional QA pairs from the SFT dataset could be used to scale up GRPO or DPO training, we restrict the dataset size in this work for two reasons: (1) the primary goal of the RL data is to analyze how SFT with LLT affects downstream RL performance, and (2) we observe that scaling beyond this size does not yield further gains. Overall, our dataset supports a wide range of post-training settings, including SFT, online RL, and offline RL.

4. Analysis on Audio Reasoning (May change title, 2, 3 written based on expected result)

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.

[**Jeff:** Your visualization figure will be here.]

Does richer perceptual information benefit audio reasoning? To evaluate how additional perceptual information affects audio reasoning, we use a strong audio–language reasoning setup consisting of detailed audio descriptions and Qwen3-32B (Yang et al., 2025) 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., 2025; 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 **??**. 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.

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. In addition, for each setting, we measure accuracy when the input audio is replaced with silence, following prior works (**?**), in order to estimate the degree of audio awareness and audio contribution.

The results in Table **??** show that the LLT pipeline successfully generates more complex and audio-dependent questions. 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 SFT data size from ?? to ??, we observe consistent performance gains when training on the mixed dataset, as shown in Figure **??**. This suggests that question complexity plays an important role in effectively scaling SFT data, even when the underlying audio content remains unchanged.

Does explicitly encoding cognitive behaviors benefit audio reasoning? To analyze the effect of reasoning behaviors, we select ?? multiple-choice questions from AF-Think and collect the corresponding ?? reasoning traces generated by LLT. We first perform SFT using these reasoning traces and then apply GRPO using the same set of questions. We compare both the resulting reasoning behaviors and RL performance across models. Following the framework proposed in (Gandhi et al., 2025), we quantify the occurrence of different cognitive behaviors in the generated reasoning traces.

Although all models are trained on identical question–answer pairs, LLT-trained models exhibit a substantially higher frequency of explicit cognitive behaviors. Moreover, these models achieve better RL performance and scale more effectively during GRPO training. We also observe that such reasoning behaviors are not naturally acquired through RL alone, even after extensive training. This highlights the importance of explicitly encoding cognitive behaviors during the SFT stage to enable more effective gains during subsequent RL optimization.

Method	MMAU test-mini	MMAU	MMAR	MMSU
Large Audio Language Models				
Audio Flamingo 3	73.3	72.4	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				
+ LLT SFT				
+ LLT SFT + DPO				
+ LLT SFT + GRPO				

Table 2. Performance comparison across different models.**[Jeff: Can you check]**

5. Experiments

5.1. Setup

Training Details. SFT detail + RL details

Baselines. Qwen2.5-omni trained on original AF-Think (same audios). Recent large-audio language models and audio reasoning models.

Evaluation detail. Widely used MMAU / MMAR / MMSU. Use their official evaluation script (string matching for MMAU / MMAR and letter matching for MMSU) for evaluation.

5.2. Results

Comparison to baselines. Outperform original AF-Think. Performance further scale up with RL. Outperform all other models. Table 2

Ablation on the LLT pipeline. We conduct ablation using 50k audios and corresponding 120k LLT reasoning pairs for the ablation. Tabel 3 Model component:

Stage 3:

Stage 4:

Generalization

general perceptual ability: wow / trea

Method	MMAU test-mini	MMAR	MMSU
Training Components			
LLM only			
+ Projector			
+ Audio Encoder			
Stage 3: Wrong CoT Traces			
No z^-			
+ z^- from M_{base}			
+ Generated z^-			
Stage 4: Reasoning Trace			
Thought expansion			
Thought continuation			

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

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

Table 4. Perforamnce across different benchmarks

Reasoning ability generlize across domain: omni-bench, mmluT

Table 4

Can be applied to different audio model: opus-lm. Maybe add qwen2-audio also?

Table 5

Qualitative Examples

count of cognitive behaviors / response length

real reasoning comparison

more qualitative examples at appendix

6. Conclusion

Concluding remarks.

Method	MMAU test-mini	MMAR	MMSU
Opus-LM			
+ LLT SFT			
Qwen2-Audio-Instruct			
+ LLT SFT			

Table 5

Impact Statement

[**Huck:** Can you write this part later..?] Authors are **required** to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. This statement should be in an unnumbered section at the end of the paper (co-located with Acknowledgements – the two may appear in either order, but both must be before References), and does not count toward the paper page limit. In many cases, where the ethical impacts and expected societal implications are those that are well established when advancing the field of Machine Learning, substantial discussion is not required, and a simple statement such as the following will suffice:

"This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here."

The above statement can be used verbatim in such cases, but we encourage authors to think about whether there is content which does warrant further discussion, as this statement will be apparent if the paper is later flagged for ethics review.

References

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, 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.

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

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.

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.

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.

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*, 2025.

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.

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.

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.

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*, 2025.

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*, 2025.

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

[**Jeff:** Add one example with some random prompt?]

B.2. Stage 2

B.3. Stage 3

B.4. Stage 4

C. More qualitative examples

D. Detailed results for each benchmark