

LongListeningThoughts: Scaling Reasoning Traces for Large Audio-Language Models

Jaeyeon Kim & Luoyi Zhang & Fernando Ruiloba Portilla
{jaeyeon2, luoyiz, fruiloba}@andrew.cmu.edu
10-423/623 Generative AI Course Project

November 25, 2025

Abstract

We propose LongListeningThoughts (LLT), a new dataset designed to improve the reasoning capabilities of large audio-language models (LALMs) through long, cognitively structured Chain-of-Thought (CoT) traces enriched with fine-grained perceptual information. Our pipeline integrates multi-level audio captions, expands existing short CoTs, and uses DeepSeek-R1 to generate long-form reasoning that incorporates cognitive behaviors such as verification and backtracking. We will fine-tune Qwen2.5-Omni on the LLT dataset and evaluate the resulting model on MMAU and MMAR, two widely used benchmarks for audio reasoning. We hypothesize that combining perceptual grounding with explicit cognitive structure will enhance both quantitative accuracy and the qualitative coherence of model reasoning.

1 Introduction

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, Ma et al., 2025b], 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 [Xu et al., 2025b] and spectrogram-based prompting. We then extend short CoTs from AF-Think [Goel et al., 2025] into long-form reasoning traces using DeepSeek-R1 [Guo et al., 2025], incorporating behaviors such as verifica-

tion and backtracking. These traces aim to provide both richer supervision during training and stronger perceptual grounding.

We will fine-tune Qwen2.5-Omni [Xu et al., 2025a] on the LLT dataset and evaluate the resulting model on MMAU [Sakshi et al., 2025] and MMAR [Ma et al., 2025b], two established benchmarks for audio reasoning. We expect LLT to improve both quantitative accuracy and the qualitative structure of model reasoning, demonstrating the value of cognitively motivated long-form CoT traces for auditory understanding.

2 Dataset / Task

We evaluate the understanding and reasoning abilities of Large Audio-Language Models (LALMs) through an audio question-answering (AQA) task. In this task, a model is presented with an audio input and a corresponding textual question, and it must generate an answer in either a closed-form (e.g., multiple-choice) or open-ended format. Formally, each datapoint consists of a triplet $\langle A, q, a \rangle$, where A is the audio signal, q is the question, and a is the ground-truth answer. Given the pair $\langle A, q \rangle$, the model is required to predict a .

For multiple-choice questions (MCQs), accuracy is computed by checking whether the model’s predicted choice matches the correct label a . Open-ended generation is more challenging to evaluate and often requires human or LLM-assisted scoring. In this work, we focus primarily on the multiple-choice setting due to its objective evaluation, scalability, and reproducibility.

Recently, several benchmarks spanning diverse audio domains, including speech, environmental sound, and music have been introduced [Sakshi et al., 2025, Ma et al., 2025b, Yang et al., 2024, Wang et al., 2025]. Among these, we select two datasets that directly target reasoning-oriented evaluation. We first use MMAU [Sakshi et al., 2025], a widely adopted benchmark that focuses on information-extraction and reasoning tasks such as temporal event reasoning and acoustic source

inference. From its full test set, we adopt the test-mini split, which contains 1,000 questions. We also include the more recently introduced MMAR [Ma et al., 2025b], which targets more challenging multi-step reasoning that requires both perceptual processing and domain-specific knowledge. Each question in MMAR is hierarchically categorized into one of four reasoning layers: Signal, Perception, Semantic, and Cultural. It also spans a diverse mixture of audio domains, including not only single-source but also multi-source mixtures such as speech–music and speech–sound combinations. MMAR provides a test set containing 1,000 questions, which we will use for our evaluation.

Both benchmarks are in MCQ format, similar to their counterparts in the large-language model (LLM) domain (e.g., MMLU [Wang et al., 2024]) and the vision-language model (VLM) domain (e.g., MMMU [Yue et al., 2024]). We report accuracy, defined as the proportion of questions for which the model predicts the correct answer choice given the audio and the question. Following the evaluation protocols of each dataset, we will apply their official string-matching procedures to extract and score the model’s predicted choice from free-form generated outputs of LALMs.

3 Related Work

Recent advances in Large Audio-Language Models (LALMs) have substantially improved performance on general audio understanding and reasoning tasks. Similar to vision-language models (VLMs), LALMs typically connect an audio encoder to an LLM and train on large-scale audio–text pairs. For example, Qwen2-Audio [Chu et al., 2024] feeds representations from a Whisper encoder [Radford et al., 2023] into the Qwen-7B LLM [Bai et al., 2023], achieving strong results across diverse audio tasks. Recent omni-models further integrate multiple modalities, such as audio, video, and text, into a unified LLM. Among them, Qwen2.5-Omni [Xu et al., 2025a] demonstrates strong performance on audio-related reasoning tasks and is widely used as a baseline for audio reasoning. In this work, we adopt Qwen2.5-Omni as our baseline.

To improve the reasoning capabilities of LALMs, several prior works have explored applying Chain-of-Thought (CoT) techniques to audio tasks. AudioCoT [Ma et al., 2025a] evaluated representative CoT strategies including few-shot CoT and zero-shot CoT. When applied to Qwen2-Audio, these approaches yielded modest improvements on MMAU. More recently, AudioReasoner [Zhifei et al., 2025] introduced CoTA, a large-scale dataset containing 1.2M structured reasoning traces across planning, captioning, and summarization tasks. Training Qwen2-Audio on CoTA led to sub-

stantial gains in audio reasoning. Audio-Thinker [Wu et al., 2025] further explored training strategies by generating CoT data from model-generated descriptions and applying reinforcement learning with rewards targeting adaptive reasoning style, accuracy, and consistency. AudioMCQ [He et al., 2025] constructed a 570k CoT dataset from LALM-generated descriptions and structured reasoning, and examined how audio contributes to performance in each dataset, while exploring different training orders based on audio contribution. Both Audio-Thinker and AudioMCQ demonstrate improvements over earlier audio reasoning methods. Compared to prior works, we focus on generating higher-quality CoT supervision by extending existing short CoTs into long-form, cognitively structured reasoning traces. We hypothesize that these richer reasoning traces lead to more robust and reliable audio reasoning performance.

Another relevant line of work explores test-time scaling to enhance reasoning in LLMs and VLMs. Increasing test-time computation via longer reasoning traces has been shown to improve performance on complex tasks such as coding and mathematics [Guo et al., 2025]. Long Perceptual Thoughts (LPT) [Liao et al., 2025] applied this idea to vision-language reasoning by constructing MCQs from dense image descriptions, generating CoT traces from VLMs, and extending them with a dedicated reasoning model (e.g., DeepSeek-R1) to introduce verification and backtracking. This approach improved performance on visual reasoning benchmarks. LongGroundedThoughts [Acuna et al., 2025] further refined this pipeline by incorporating bounding-box grounding and scaling training strategies. Motivated by these findings, we investigate whether long, cognitively structured reasoning traces can similarly enhance perceptual reasoning in the audio domain. To this end, we propose a new pipeline for constructing such reasoning datasets tailored specifically for auditory understanding.

4 Approach

Baseline. As a baseline, we evaluate an existing LALM, Qwen2.5-Omni, by directly running it on MMAU and MMAR without additional training, as it already demonstrates strong auditory understanding and reasoning ability. As an additional baseline, we perform supervised fine-tuning on the original short CoT dataset used in our method and compare its performance against the model trained with our proposed dataset, contingent on available computational resources.

LongListeningThoughts. We propose LongListeningThoughts (LLT), a new dataset generation pipeline

Name	MMAU (Test-mini)				MMAR			
	Sound	Music	Speech	Avg	Sound	Music	Speech	Avg
Qwen2.5-omni-7b	69.67	64.37	63.66	65.60	58.79	40.78	59.86	56.70
Qwen2.5-omni-7b (reproduced)	81.68	68.26	73.87	74.60	64.24	50.49	65.99	61.5
Qwen2.5-omni-7b + SFT on AF-Think								
Qwen2.5-omni-7b + SFT on LLT								

Table 1: Accuracy (%) of models on MMAU (test-mini) and MMAR across audio domains.

designed to produce long-form, cognitively structured reasoning traces. These traces explicitly model cognitive behaviors such as verification, backtracking, and subgoal setting, while enhancing perceptual grounding through the integration of multi-level detailed audio descriptions. Unlike existing methods that build CoT datasets from scratch, LLT operates by extending and enriching existing short CoT traces, making it modular, scalable and compatible with prior approaches [Zhifei et al., 2025, Goel et al., 2025, He et al., 2025].

Our pipeline begins by enriching each example with dense audio descriptions that provide fine-grained perceptual information to the reasoning model. To capture semantic and event-level information, such as overall temporal structure, relationships among sound events, we use Qwen3-Omni-Captioner [Xu et al., 2025b] to generate audio captions. To capture lower-level acoustic attributes, such as duration, timbre, amplitude, frequency, reverberation, and spectral characteristics, we prompt Qwen3-VL-32B [Team, 2025] to describe linear spectrograms of the corresponding audio clips. These complementary descriptions provide rich perceptual signals that guide the generation of more grounded long-form reasoning traces.

We then augment existing CoT datasets using these dense descriptions. Starting from AF-Think [Goel et al., 2025], which contains approximately 250k short CoT examples spanning diverse audio skills, we select a subset of 20k examples as candidates for extension. To model both correct and corrective reasoning, we retain half of the examples as-is and generate incorrect short CoTs for the remaining half by instructing an LLM to produce reasoning chains that lead to wrong answers. Let z^+ and z^- denote correct and incorrect short CoTs, respectively, paired with corresponding answers a^+ and a^- . This gives us both $\langle z^+, a^+ \rangle$ pairs directly from AF-Think and $\langle z^-, a^- \rangle$ pairs derived from the augmented set.

Next, we use these short CoTs and dense audio descriptions to generate long, cognitively structured reasoning traces using DeepSeek-R1 [Guo et al., 2025]. Following the approach of LongPerceptualThoughts [Liao et al., 2025], we prompt the model with (1) the short CoT and (2) the dense audio descriptions, instructing it to complete the reasoning process. To en-

courage richer cognitive behaviors, we additionally insert trigger cues such as “wait” and “let me check,” which induce the model to exhibit reasoning behaviors such as verification and backtracking. This process yields long-form reasoning traces z_ℓ , which we pair with both correct and incorrect short CoTs, producing datasets of the form $\langle z^+, z_\ell, a^+ \rangle$ and $\langle z^-, z_\ell, a^+ \rangle$.

Finally, we perform supervised fine-tuning of an LALM using three types of training pairs: (1) correct short CoT \rightarrow answer $\langle z^+, a^+ \rangle$, (2) correct short CoT + long CoT \rightarrow answer $\langle z^+, z_\ell, a^+ \rangle$, and (3) incorrect short CoT + long CoT \rightarrow answer $\langle z^-, z_\ell, a^+ \rangle$. We construct 30k training examples, 10k for each category.

5 Experiment

Experiment Details. We evaluate our approach on 1,000 samples from MMAU Test-mini [Sakshi et al., 2025] and 1,000 samples from MMAR [Ma et al., 2025b]. As our baseline model, we use Qwen2.5-Omni-7B [Xu et al., 2025a], which we evaluate directly without additional training. To assess the effectiveness of LLT, we perform supervised fine-tuning (SFT) on 30k LLT examples using LlamaFactory [Zheng et al., 2024]. For comparison, we also train a secondary baseline by performing SFT on the original 20k AF-Think short CoT dataset, the same dataset from which LLT is constructed, allowing us to measure the added benefit of our long-form reasoning augmentation.

Results. Table 1 presents the overall accuracy of each model across the sound domains for both MMAU and MMAR. The skeleton table includes columns for performance on each domain as well as the averaged accuracy. It reports baseline performance of Qwen2.5-Omni-7B and reproduction of it, and performance of the short-CoT-fine-tuned model, performance of the LLT-trained mode

Our preliminary reproduction of the baseline Qwen2.5-Omni-7B results shows better than expected performance at around 74% accuracy on MMAU Test-mini and 61.5% accuracy on MMAR. These results are around 9% and 5% higher than previously reported trends for models of similar scale. We expect the LLT-trained model to outperform both the vanilla baseline and the short-CoT-fine-tuned model, reflecting the

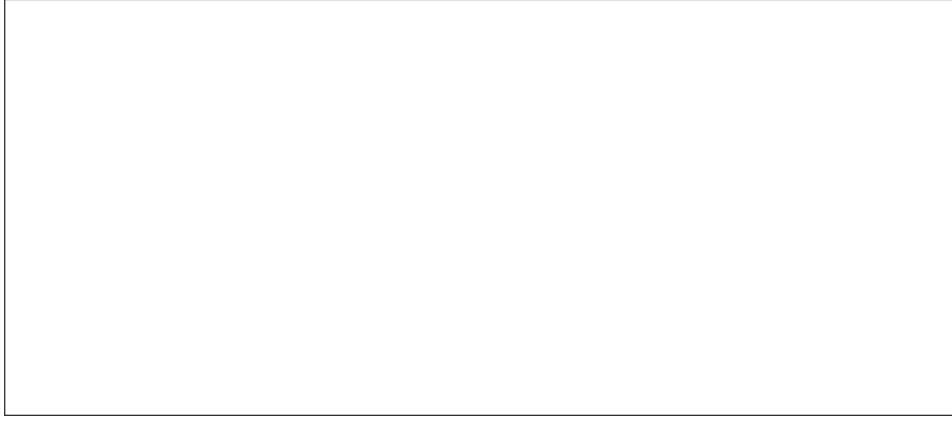


Figure 1: Qualitative results. (a) Examples comparing model behavior between Qwen2.5-Omni and the LLT-trained model. (b) Aggregate counts of reasoning behaviors such as verification, backtracking, and subgoal-setting.

benefit of long-form, cognitively structured reasoning traces.

Qualitative Analysis. Additionally, we will perform qualitative analysis to examine how LLT affects reasoning behaviors. Figure 1 provides the skeleton for this analysis. The left portion of the figure will show example model outputs for the same question under two conditions: (1) the vanilla Qwen2.5-Omni-7B model, and (2) the LLT-trained model. We will highlight segments in the model’s responses that correspond to reasoning behaviors such as verification or backtracking, illustrating how LLT promotes richer cognitive structure.

The right portion of Figure 1 will present aggregate statistics of these behaviors. We will generate a fixed number of model outputs from each system and use GPT-5.1 as an automatic evaluator to count occurrences of reasoning behaviors such as “subgoal-setting,” “verifying,” or “backtracking.” This analysis will allow us to measure whether LLT encourages more explicit cognitive reasoning patterns.

6 Plan

By November 29, we plan to finalize the dataset generation pipeline and produce the full set of 30k LLT training examples, with Jaeyeon leading this task. By December 6, we will complete supervised fine-tuning on the generated dataset, for which Luoyi will take primary responsibility. Following training, we will run evaluations on MMAU and MMAR by December 8, with Fernando overseeing the evaluation process. Finally, all team members will collaborate to complete the final presentation and written report. Throughout the project, the team will continue to support one another across tasks as needed.

7 Thought-Experiment on Compute

In our actual workflow, we relied on compute resources available through Jaeyeon’s PhD program. Specifically, we used a combination of NVIDIA L40s H100 GPUs. Although we do not track exact GPU hours, we estimate that dataset generation required approximately 1 day on L40s GPUs, while each model training required roughly 6 hours on 4 H100 GPUs. If these computations were performed on commercial cloud platforms such as AWS, the equivalent cost would be substantial. For example, 8* H100 instances typically cost \$50 per GPU hour, and L40s instances approximately \$2 per GPU hour. But these computations were conducted on institutional clusters, they incurred no direct monetary cost to our project. If an additional \$450 in cloud GPU credits were available, our overall experimental design would remain largely unchanged since the project is not primarily constrained by cloud-compute cost.

References

- David Acuna, Chao-Han Huck Yang, Yuntian Deng, Jaehun Jung, Ximing Lu, Prithviraj Ammanabrolu, Hyunwoo Kim, Yuan-Hong Liao, and Yejin Choi. Long grounded thoughts: Distilling compositional visual reasoning chains at scale. *arXiv preprint arXiv:2511.05705*, 2025.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng

He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, et al. Measuring audio’s impact on correctness: Audio-contribution-aware post-training of large audio language models. *arXiv preprint arXiv:2509.21060*, 2025.

Yuan-Hong Liao, Sven Elflein, Liu He, Laura Leal-Taixé, Yejin Choi, Sanja Fidler, and David Acuna. Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception. *arXiv preprint arXiv:2504.15362*, 2025.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*, 2025a.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025b.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

Qwen Team. Qwen3-vl: Sharper vision, deeper thought, broader action. <https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8ad9fef>, 2025. Accessed: 2025-11-24.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. Mmlupro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

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

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*, 2025.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruofu Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

- 434 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yan-
435 han Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang
436 Ma. Llamafactory: Unified efficient fine-tuning
437 of 100+ language models. In *Proceedings of the*
438 *62nd Annual Meeting of the Association for Com-*
439 *putational Linguistics (Volume 3: System Demon-*
440 *strations)*, Bangkok, Thailand, 2024. Association
441 for Computational Linguistics. URL [http://](http://arxiv.org/abs/2403.13372)
442 arxiv.org/abs/2403.13372.
- 443 Xie Zhifei, Mingbao Lin, Zihang Liu, Pengcheng
444 Wu, Shuicheng Yan, and Chunyan Miao. Audio-
445 reasoner: Improving reasoning capability in large
446 audio language models. In *Proceedings of the 2025*
447 *Conference on Empirical Methods in Natural Lan-*
448 *guage Processing*, pages 23840–23862, 2025.