

# 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

## 1 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

77 inference. From its full test set, we adopt the test-  
78 mini split, which contains 1,000 questions. We also in-  
79 clude the more recently introduced MMAR [Ma et al.,  
80 2025b], which targets more challenging multi-step rea-  
81 soning that requires both perceptual processing and  
82 domain-specific knowledge. Each question in MMAR  
83 is hierarchically categorized into one of four reasoning  
84 layers: Signal, Perception, Semantic, and Cultural. It  
85 also spans a diverse mixture of audio domains, includ-  
86 ing not only single-source but also multi-source mix-  
87 tures such as speech–music and speech–sound combi-  
88 nations. MMAR provides a test set containing 1,000  
89 questions, which we will use for our evaluation.

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

### 101 3 Related Work

102 Recent advances in Large Audio-Language Models  
103 (LALMs) have substantially improved performance on  
104 general audio understanding and reasoning tasks. Sim-  
105 ilar to vision-language models (VLMs), LALMs typ-  
106 ically connect an audio encoder to an LLM and train  
107 on large-scale audio–text pairs. For example, Qwen2-  
108 Audio [Chu et al., 2024] feeds representations from a  
109 Whisper encoder [Radford et al., 2023] into the Qwen-  
110 7B LLM [Bai et al., 2023], achieving strong results  
111 across diverse audio tasks. Recent omni-models further  
112 integrate multiple modalities, such as audio, video, and  
113 text, into a unified LLM. Among them, Qwen2.5-Omni  
114 [Xu et al., 2025a] demonstrates strong performance on  
115 audio-related reasoning tasks and is widely used as a  
116 baseline for audio reasoning . In this work, we adopt  
117 Qwen2.5-Omni as our baseline.

118 To improve the reasoning capabilities of LALMs,  
119 several prior works have explored applying Chain-of-  
120 Thought (CoT) techniques to audio tasks. AudioCoT  
121 [Ma et al., 2025a] evaluated representative CoT strate-  
122 gies including few-shot CoT and zero-shot CoT. When  
123 applied to Qwen2-Audio, these approaches yielded  
124 modest improvements on MMAU. More recently, Au-  
125 dioReasoner [Zhifei et al., 2025] introduced CoTA, a  
126 large-scale dataset containing 1.2M structured reason-  
127 ing traces across planning, captioning, and summariza-  
128 tion tasks. Training Qwen2-Audio on CoTA led to sub-

stantial gains in audio reasoning. Audio-Thinker [Wu  
129 et al., 2025] further explored training strategies by gen-  
130 erating CoT data from model-generated descriptions  
131 and applying reinforcement learning with rewards tar-  
132 geting adaptive reasoning style, accuracy, and consis-  
133 tency. AudioMCQ [He et al., 2025] constructed a 570k  
134 CoT dataset from LALM-generated descriptions and  
135 structured reasoning, and examined how audio con-  
136 tributes to performance in each dataset, while explor-  
137 ing different training orders based on audio contribu-  
138 tion. Both Audio-Thinker and AudioMCQ demon-  
139 strate improvements over earlier audio reasoning meth-  
140 ods. Compared to prior works, we focus on generating  
141 higher-quality CoT supervision by extending existing  
142 short CoTs into long-form, cognitively structured rea-  
143 soning traces. We hypothesize that these richer reason-  
144 ing traces lead to more robust and reliable audio rea-  
145 soning performance.

146 Another relevant line of work explores test-time  
147 scaling to enhance reasoning in LLMs and VLMs. In-  
148 creasing test-time computation via longer reasoning  
149 traces has been shown to improve performance on  
150 complex tasks such as coding and mathematics [Guo  
151 et al., 2025]. Long Perceptual Thoughts (LPT) [Liao  
152 et al., 2025] applied this idea to vision-language rea-  
153 soning by constructing MCQs from dense image de-  
154 scriptions, generating CoT traces from VLMs, and ex-  
155 tends them with a dedicated reasoning model (e.g.,  
156 DeepSeek-R1) to introduce verification and backtrack-  
157 ing. This approach improved performance on vi-  
158 sual reasoning benchmarks. LongGroundedThoughts  
159 [Acuna et al., 2025] further refined this pipeline by in-  
160 corporating bounding-box grounding and scaling train-  
161 ing strategies. Motivated by these findings, we in-  
162 vestigate whether long, cognitively structured reason-  
163 ing traces can similarly enhance perceptual reasoning  
164 in the audio domain. To this end, we propose a new  
165 pipeline for constructing such reasoning datasets tai-  
166 lored specifically for auditory understanding.

### 168 4 Approach

**Baseline.** As a baseline, we evaluate an existing  
169 LALM, Qwen2.5-Omni, by directly running it on  
170 MMAU and MMAR without additional training, as  
171 it already demonstrates strong auditory under-  
172 standing and reasoning ability. As an additional base-  
173 line, we perform supervised fine-tuning on the origi-  
174 nal short CoT dataset used in our method and compare  
175 its performance against the model trained with our pro-  
176 posed dataset, contingent on available computational  
177 resources.

**LongListeningThoughts.** We propose LongListen-  
178 ingThoughts (LLT), a new dataset generation pipeline  
179

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  $z_\ell$ , 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 → answer  $\langle z^+, a^+ \rangle$ , (2) correct short CoT + long CoT → answer  $\langle z^+, z_\ell, a^+ \rangle$ , and (3) incorrect short CoT + long CoT → 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.

269 benefit of long-form, cognitively structured reasoning  
270 traces.

271 **Qualitative Analysis.** Additionally, we will perform  
272 qualitative analysis to examine how LLT affects rea-  
273 soning behaviors. Figure 1 provides the skeleton for  
274 this analysis. The left portion of the figure will show  
275 example model outputs for the same question under  
276 two conditions: (1) the vanilla Qwen2.5-Omni-7B  
277 model, and (2) the LLT-trained model. We will high-  
278 light segments in the model’s responses that correspond  
279 to reasoning behaviors such as verification or back-  
280 tracking, illustrating how LLT promotes richer cogni-  
281 tive structure.

282 The right portion of Figure 1 will present aggregate  
283 statistics of these behaviors. We will generate a fixed  
284 number of model outputs from each system and use  
285 GPT-5.1 as an automatic evaluator to count occurrences  
286 of reasoning behaviors such as “subgoal-setting,” “ver-  
287 ifying.” or “backtracking.” This analysis will allow us  
288 to measure whether LLT encourages more explicit cog-  
289 nitive reasoning patterns.

## 7 Thought-Experiment on Compute

In our actual workflow, we relied on compute resources  
304 available through Jaeyeon’s PhD program. Specifi-  
305 cally, we used a combination of NVIDIA L40s H100  
306 GPUs. Although we do not track exact GPU hours,  
307 we estimate that dataset generation required approxi-  
308 mately 1 day on L40s GPUs, while each model train-  
309 ing required roughly 6 hours on 4 H100 GPUs. If these  
310 computations were performed on commercial cloud  
311 platforms such as AWS, the equivalent cost would be  
312 substantial. For example, 8\* H100 instances typically  
313 cost \$50 per GPU hour, and L40s instances approxi-  
314 mately \$2 per GPU hour. But these computations were  
315 conducted on institutional clusters, they incurred no  
316 direct monetary cost to our project. If an additional  
317 \$450 in cloud GPU credits were available, our overall  
318 experimental design would remain largely unchanged  
319 since the project is not primarily constrained by cloud-  
320 compute cost.  
321

## 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

## 6 Plan

291 By November 29, we plan to finalize the dataset gen-  
292 eration pipeline and produce the full set of 30k LLT  
293 training examples, with Jaeyeon leading this task. By  
294 December 6, we will complete supervised fine-tuning  
295 on the generated dataset, for which Luoyi will take pri-  
296 mary responsibility. Following training, we will run  
297 evaluations on MMAU and MMAR by December 8,  
298 with Fernando overseeing the evaluation process. Fi-  
299 nally, all team members will collaborate to complete  
300 the final presentation and written report. Throughout  
301 the project, the team will continue to support one an-  
302 other across tasks as needed.

- 335 He, Junyang Lin, et al. Qwen2-audio technical re- 336 port. *arXiv preprint arXiv:2407.10759*, 2024. 337 99f0335c4ad9ff6153e517418d48535ab6d8aefef, 338 2025. Accessed: 2025-11-24. 339 385
- 340 Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Ku- 341 mar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck 342 Yang, Ramani Duraiswami, Dinesh Manocha, 343 Rafael Valle, et al. Audio flamingo 3: Advancing au- 344 dio intelligence with fully open large audio language 345 models. *arXiv preprint arXiv:2507.08128*, 2025. 346 347
- 348 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, 349 Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, 350 Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentiviz- 351 ing reasoning capability in llms via reinforcement 352 learning. *arXiv preprint arXiv:2501.12948*, 2025. 353
- 354 Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia 355 Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi 356 Liu, Zining Liang, et al. Measuring audio’s im- 357 pact on correctness: Audio-contribution-aware post- 358 training of large audio language models. *arXiv preprint 359 arXiv:2509.21060*, 2025.
- 360 Yuan-Hong Liao, Sven Elflein, Liu He, Laura Leal- 361 Taixé, Yejin Choi, Sanja Fidler, and David Acuna. 362 Longperceptualthoughts: Distilling system-2 rea- 363 soning for system-1 perception. *arXiv preprint 364 arXiv:2504.15362*, 2025.
- 365 Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong 366 Chng, and Xie Chen. Audio-cot: Exploring chain- 367 of-thought reasoning in large audio language model. 368 *arXiv preprint arXiv:2501.07246*, 2025a.
- 369 Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, 370 Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe 371 Chen, Zhuo Chen, Jian Cong, et al. Mmar: 372 A challenging benchmark for deep reasoning in 373 speech, audio, music, and their mix. *arXiv preprint 374 arXiv:2505.13032*, 2025b.
- 375 Alec Radford, Jong Wook Kim, Tao Xu, Greg Brock- 376 man, Christine McLeavey, and Ilya Sutskever. Ro- 377 bust speech recognition via large-scale weak su- 378 pervision. In *International conference on machine 379 learning*, pages 28492–28518. PMLR, 2023.
- 380 S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, 381 Ramaneswaran Selvakumar, Oriol Nieto, Ramani 382 Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 383 Mmau: A massive multi-task audio understanding 384 and reasoning benchmark. In *The Thirteenth Inter- 385 national Conference on Learning Representations*, 386 2025.
- 387 Qwen Team. Qwen3-vl: Sharper vi- 388 sion, deeper thought, broader ac- 389 tion. <https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8aefef>, 390 2024. Accessed: 2025-11-24. 391
- 392 Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuo- 393 han Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, 394 and Nancy Chen. Audiobench: A universal bench- 395 mark for audio large language models. In *Proceed- 396 ings of the 2025 Conference of the Nations of the 397 Americas Chapter of the Association for Compu- 398 tational Linguistics: Human Language Technologies 399 (Volume 1: Long Papers)*, pages 4297–4316, 2025. 393
- 400 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, 401 Abhranil Chandra, Shiguang Guo, Weiming Ren, 402 Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu- 403 pro: A more robust and challenging multi-task 404 language understanding benchmark. *Advances in 405 Neural Information Processing Systems*, 37:95266– 406 95290, 2024.
- 407 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten 408 Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, 409 et al. Chain-of-thought prompting elicits reasoning 410 in large language models. In *Advances in neural 411 information processing systems*, volume 35, pages 412 24824–24837, 2022.
- 413 Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, 414 Hualei Wang, Meng Yu, and Dong Yu. Audio- 415 thinker: Guiding audio language model when and 416 how to think via reinforcement learning. *arXiv 417 preprint arXiv:2508.08039*, 2025.
- 418 Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting 419 He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, 420 Kai Dang, et al. Qwen2. 5-omni technical report. 421 *arXiv preprint arXiv:2503.20215*, 2025a.
- 422 Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong 423 Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting 424 He, Xinfu Zhu, et al. Qwen3-omni technical report. 425 *arXiv preprint arXiv:2509.17765*, 2025b.
- 426 Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue 427 Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, 428 Zhou Zhao, Chang Zhou, et al. Air-bench: Bench- 429 marking large audio-language models via generative 430 comprehension. *arXiv preprint arXiv:2402.07729*, 431 2024.
- 432 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, 433 Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu 434 Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: 435 A massive multi-discipline multimodal under- 436 standing and reasoning benchmark for expert agi. In *Pro- 437 ceedings of the IEEE/CVF Conference on Computer 438 Vision and Pattern Recognition*, pages 9556–9567, 439 2024.

434 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yan-  
435 han Ye, Zheyuan 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://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.