Audio-Reasoner: Improving Reasoning Capability in Large Audio Language Models

Zhifei Xie[1*] **Mingbao Lin**[3*] **Zihang Liu**[2*]
Pengcheng Wu[1] **Shuicheng Yan**[2†] **Chunyan Miao**[1†]
[1]Nanyang Technological University [2]National University of Singapore [3]Rakuten Singapore
zhifei001@e.ntu.edu.sg limb001@outlook.com liuzihang99@gmail.com
pengchengwu@ntu.edu.sg yansc@nus.edu.sg ascymiao@ntu.edu.sg

Abstract

Recent advancements in multimodal reasoning overlook the audio modality. We introduce Audio-Reasoner, a large-scale audio language model for deep reasoning. We meticulously curated a large-scale and diverse multi-task audio dataset with simple annotations. Then, we leverage closed-source models to conduct secondary labeling, QA generation, along with structured COT process. These datasets together form a high-quality reasoning dataset with 1.2 million reasoning-rich samples, which we name CoTA. Following inference scaling principles, we train Audio-Reasoner on CoTA, enabling it to achieve great logical capabilities in audio reasoning. Experiments show state-of-the-art performance across key benchmarks, including MMAU-mini (+25.42%), AIR-Bench chat/foundation (+14.57%/+10.13%), and MELD (+8.01%). Our findings stress the core of structured CoT training in advancing audio reasoning. The model, dataset, and code are open-sourced at `https://github.com/xzf-thu/Audio-Reasoner` or `https://huggingface.co/datasets/zhifeixie/Audio-Reasoner-CoTA`.

1 Introduction

Large language models (LLMs) improves reasoning by chain-of-thought (CoT) and inference scaling. OpenAI's o1 (Jaech et al., 2024) and Deepseek-R1 (Guo et al., 2025) have shown strong performance in mathematics and coding (Team et al., 2025; Zhao et al., 2024a; Muennighoff et al., 2025; Liu et al., 2024a; Zhang et al., 2024b; Yang et al., 2024a). Methods like Visual-CoT (Shao et al., 2024), LLaVA-Reasoner (Zhang et al., 2024a), and MAmmoTH-VL (Guo et al., 2024) show the benefits of combining large-scale data with multimodal reasoning (Zou et al., 2023). Mulberry (Wen et al., 2019) and Image-of-Thought (Zhou et al., 2024), incorporate reflection and image editing to further improve multimodal understanding.

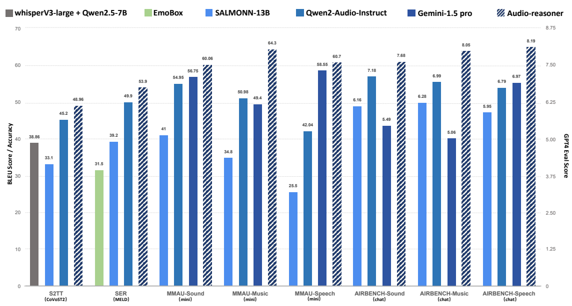

Figure 1: Benchmark performance of Audio-Reasoner. Best view with zooming in.

Although models like Audio Flamingo (Kong et al., 2024), SALMONN (Tang et al., 2023), and Qwen2-Audio (Chu et al., 2024) have pushed the boundaries of large audio language models (LALMs), these advancements have not yet incorporated CoT reasoning at scale. Recent research (Ma et al., 2025a) suggests that existing CoT methods, such as zero-shot reasoning in audio tasks, fail to significantly improve performance on more complex queries. This limitation is largely attributed to the simplicity of existing audio datasets—such as AudioSet (Gemmeke et al., 2017), AudioCaps (Kim et al., 2019), and Clotho (Drossos et al., 2020)—which predominantly feature short, simple labels. These simplified datasets hinder the development of LALMs capable of more intricate reasoning. Without richer, more complex data, these models struggle with long-form reasoning, and the application of CoT often leads to degraded performance.

To address the challenges in audio-based reasoning, we propose a scalable and effective approach to generating high-quality pretraining data. We introduce CoTA, a large-scale dataset containing **1.2 million** refined captions and question-answer (QA) pairs. CoTA spans multiple datasets and

*Contributed equally.

†Shuicheng Yan and Chunyan Miao are co-corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23841–23863
November 4-9, 2025 ©2025 Association for Computational Linguistics

tasks, undergoing rigorous filtering to ensure diversity and quality. Building on CoTA, we develop Audio-Reasoner, a LALM for long-context reasoning. Audio-Reasoner is trained with a 4K token context window and generates structured CoT reasoning with length could more than exceeding 1K tokens in real-world tasks. The model is fine-tuned on CoTA, adhering to a structured reasoning framework, as illustrated in Figure 2: (1) Planning—Identifies key problem components from the user query and formulates a structured sequence of reasoning steps essential for deriving an answer. (2) Caption—Extracts and integrates relevant multimodal content from the input to enrich the reasoning process. (3) Reasoning— Executes a systematic, step-by-step reasoning procedure to ensure logical coherence and accuracy. (4) Summary—Synthesizes the reasoning process into a final response that is concise, grounded, and precise.

Our experimental results, partially presented in Figure 1, demonstrate the effectiveness of Audio-Reasoner. We evaluate the model across multiple benchmarks: MMAU-mini (Sakshi et al., 2024): A dataset with 1,500 closed-choice questions testing reasoning across sound, speech, and music. AIR-Bench (Yang et al., 2024b): Various types of audio signals including human speech, natural sounds, and music. CoVoST2(zh-en) (Wang et al., 2021): Speech-to-text translation in Chinese and English. MELD (Poria et al., 2019): Emotion classification. Compared to Qwen2-Audio-Instruct (Chu et al., 2024), Audio-Reasoner achieves: +25.4% improvement on MMAU-mini with reasoning subtask gains: +44.4%, +26.1%, and +9.3%; +14.6% gains on AIR-Bench chat; +30.6% on CoVoST2(ZN/EN subset, Average BLEU score.); +8.01% on MELD.

The major contributions we have made include:

1. We propose Audio-Reasoner for deep audio reasoning and inference scalingy. Built upon Qwen2-Audio and fine-tuned with structured CoT training, Audio-Reasoner improves long-context reasoning across diverse audio tasks.
2. We develop CoTA, a large-scale dataset with 1.2 million high-quality captions and QA pairs, spanning multiple audio domains. The dataset enables structured reasoning and enhances audio-language pretraining.
3. We introduce a scalable data generation pipeline leveraging commercial models to produce complex reasoning-based QA pairs and structured CoT annotations, enriching model training.
4. We achieve best results, with +25.4% gains over Qwen2-Audio-Instruct on MMAU-mini, along with significant improvements in reasoning, translation, and emotion recognition tasks.

2 Related Work

Chain-of-Thought Reasoning. LLMs enhance reasoning through in-context learning (ICL), and is improved by CoT. CoT has been explored in various forms, including Tree of Thoughts (Yao et al., 2023), manual few-shot prompts (Wei et al., 2022), and automatically generated methods (Jin et al., 2024; Zhang et al., 2022). Recent studies have examined the necessity, theory, and broad applicability of CoT (Sprague et al., 2024; Stechly et al., 2024; Turpin et al., 2023). OpenAI's o1 model (Jaech et al., 2024) has reignited interest in CoT, showing strong performance in complex tasks like coding (Zhang et al., 2024b) and math (Yang et al., 2024a). CoT is often combined with strategies like Monte Carlo Tree Search (Browne et al., 2012), reflection (Guo et al., 2025), tool use (Qin et al., 2023), and reinforcement learning (Rafailov et al., 2023) to boost performance.

Multimodal Chain-of-Thought. CoT are also explored in multimodal models. LLaVA-Reasoner (Zhang et al., 2024a) and LLaVA-CoT (Xu et al., 2024) leverage recaptioning and dataset scaling for CoT fine-tuning. MAmmoTH-VL (Guo et al., 2024) improves performance through large-scale training. Mulberry (Wen et al., 2019) and Image-of-Thought (Zhou et al., 2024) enhance reasoning using reflection and image editing tools. In videos, several studies (Wang et al., 2024; Han et al., 2024; Fei et al., 2024; Tang et al., 2024) show CoT's effectiveness. But CoT in audio remains underdeveloped–Audio-COT (Ma et al., 2025a) shows limited gains for complex tasks.

Large Audio Language Models. LALMs include: audio understanding and real-time dialogue. The former typically consists of a three-layer architecture—an encoder, connector, and an LLM—focusing on specific domains, as seen in models like Mu-LLaMA (Liu et al., 2024b), LTU (Gong et al., 2023b), EmoBox (Ma et al., 2024), and GAMA (Ghosh et al., 2024). Other models, such as LTU-AS (Gong et al., 2023a), SALMONN (Tang et al., 2023) and Qwen2-Audio (Chu et al., 2024), employ unified architectures designed for multi-task training. The latter, which focuses on speech input and extend transformers to real-time speech synthesis, is also gaining popularity (Zhang et al., 2023a; Xie and Wu, 2024a,b; Fu et al., 2025; Défossez et al., 2024). However, current LALMs still lack significant exploration into reasoning.

3 Audio-Reasoner

3.1 Model Training with Audio Reasoning

A standard large language model is trained to generate an output sequence Y given an input sequence X. The probability distribution of the model's output is formulated as:

$$P(Y|X;\theta) = f_\theta(X), \tag{1}$$

where f_θ is a Transformer-based model parameterized by θ. The training objective follows a maximum likelihood estimation framework:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log P(Y_i|X_i;\theta). \tag{2}$$

In our Audio-Reasoner, the input consists of an audio signal A and a text-based query Q, forming the multimodal input representation:

$$X = (A, Q). \tag{3}$$

Unlike conventional LLMs, where the output is a single response, we structure the model's output into two distinct components: the chain of thought reasoning C, which captures the step-by-step logical process, and the final response R, which provides the ultimate answer. The model thus learns to generate the concatenation of C and R, leading to the probability distribution:

$$P(C,R|A,Q;\theta) = f_\theta(A,Q). \tag{4}$$

To ensure explicit learning of both reasoning and final response generation, we construct a dataset:

$$\mathcal{D} = \{(A_i, Q_i, C_i, R_i)\}_{i=1}^{N}, \tag{5}$$

where each training sample consists of an input audio signal A_i, its corresponding textual query Q_i, the structured reasoning process C_i, and the final answer R_i. This dataset formulation reinforces the model's ability to perform in-context learning and deep reasoning, ensuring that generated responses are not only accurate but also logically structured.

The training objective maximizes the likelihood of both C and R, encouraging the model to first reason and then generate a response.

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log P(C_i, R_i|A_i, Q_i;\theta). \tag{6}$$

Optimizing this objective trains Audio-Reasoner to generate structured reasoning before the final answer, improving interpretability, reliability, and alignment with human thinking.

At inference time, Audio-Reasoner follows a structured reasoning in Figure 2. (1) **Planning** (P): The model analyzes the query, identifies key problem components, and outlines the reasoning steps necessary to derive an answer. (2) **Captioning** (C): Relevant multimodal content is extracted from the input, such as speech transcription, acoustic event detection, or context information. (3) **Reasoning** (R): Based on the extracted content, the model performs structured, step-by-step reasoning. (4) **Summary** (S): The model synthesizes its reasoning process into a concise and accurate response.

$$P \sim f_\theta(A,Q), \tag{7}$$
$$C \sim f_\theta(A,Q,P), \tag{8}$$
$$R \sim f_\theta(A,Q,P,C), \tag{9}$$
$$S \sim f_\theta(A,Q,P,C,R). \tag{10}$$

Compared with the direct-response method (Chu et al., 2024), this approach provides two key advantages: **Improved Interpretability**—By modeling each reasoning step, the process becomes more transparent, making it easier to analyze and diagnose errors. **Reduced Hallucinations**—The structured reasoning mitigates speculative or incorrect responses, ensuring that outputs remain grounded.

Figure 2 shows the structured CoT reasoning process. It draws inspiration from recent symbolic reasoning and CoT training (Cui et al., 2024), which emphasize zero-shot reasoning without training is less effective. Also, previous studies have shown that models tuned on native CoT data surpass those trained on generic labels, especially in multimodal reasoning tasks (Guo et al., 2024; Wen et al., 2019).

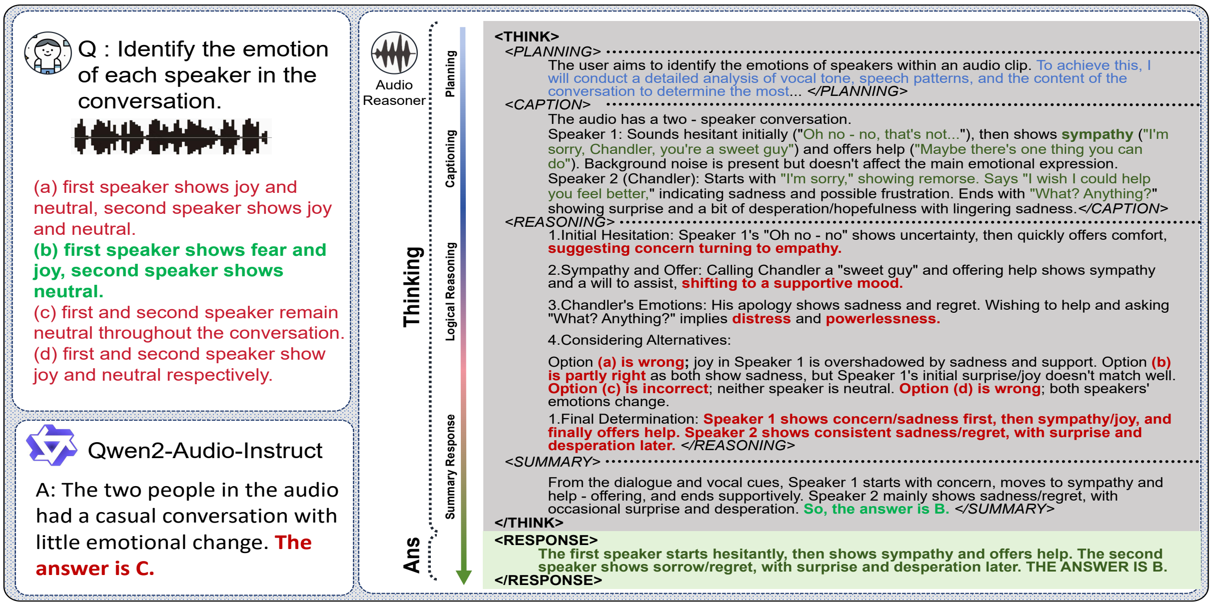

Figure 2: Comparison between Audio-Reasoner and Qwen2-Audio-Instruct: While Qwen2-Audio-Instruct produces brief and error-prone reasoning, our Audio-Reasoner uses a structured reasoning framework with distinct "thinking" and "response" phases, ensuring a more reliable and well-supported output through careful planning, information extraction, and step-by-step reasoning.

3.2 Systematic Data Preparation

Training the Audio-Reasoner model requires a high-quality, diverse, and multitask audio-based reasoning dataset. Our goal is to develop a scalable and effective data generation method that systematically transforms raw audio data and simple human-labeled annotations into structured reasoning tasks. The resulting CoTA dataset with 1.2 million samples, focusing on complex reasoning-based question-answering tasks, spans three domains of audio, speech, and music.

To achieve this, our structured data generation pipeline includes: (1) generating high-quality annotations and diverse questions, (2) constructing structured reasoning chains, and (3) performing comprehensive validation. The complete pipeline is illustrated in Figure 3.

3.2.1 Multistage Data Generation Pipeline

Stage 1: Automated Annotation and Question-Answer Synthesis. We first use advanced closed-source models to enhance basic human annotations into high-quality, coherent training data. While large language models may hallucinate in free-form tasks, they excel at structured, evidence-based generation. We guide the model to describe audio elements in sequence, helping it understand sound sources and speech context. From these enriched descriptions, the model generates a wide range of questions—from simple factual ones to complex reasoning tasks—ensuring broad coverage of reasoning types. The detailed prompting strategy used in this stage is provided in Sec. A.1.

Stage 2: Structured Reasoning Chain Construction. Next, we transform the generated question-answer pairs into structured reasoning chains. Given the limited development of CoT methodologies in the audio domain, we adopt a systematic approach to ensure inference stability. The model first plans and analyzes the questions, extracts key information from the captions, and formulates logical steps leading to the answer. To facilitate structured reasoning, we employ explicit step demarcations such as <THINK> and <REASONING> construct multi-step inference pathways. Sec. A.2 describes the prompt for the structured reasoning chain construction process.

Stage 3: Quality Assurance and Dataset Validation. Finally, we subject the generated data to a rigorous review process. Using the raw audio input, Stage 1 annotations, and Stage 2 reasoning chains, the model assesses whether the generated content is accurate, coherent, and suitable for inclusion in the final dataset. This step ensures the overall quality and reliability of the CoTA dataset. Sec. A.3 shows prompt used for filtering low-quality contents.

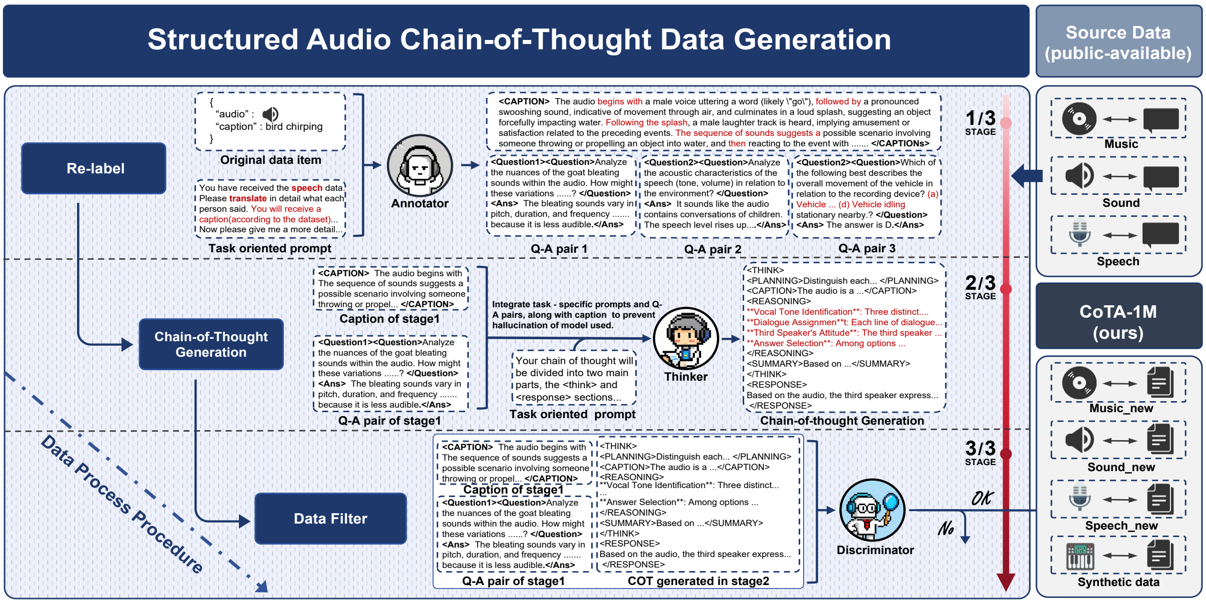

Figure 3: Multistage data generation pipeline.

3.2.2 Task Taxonomy

CoTA encompasses a range of reasoning-based tasks, each requiring distinct reasoning path that the model should grasp. These include:

(1) Sound-Based Question Answering: The model identifies and analyzes sound characteristics, contextualizing them within the user's query to derive a reasoned response. **(2) Speech-Based Question Answering**: The model recognizes speaker timbres, transcribes speech content, and incrementally processes the question to determine the appropriate answer. **(3) Speech Emotion Recognition (SER) and Speech-to-Text Translation (S2TT)**: These specialized tasks require the model to integrate speech recognition with emotion analysis and language translation, forming a structured reasoning process. **(4) Music-Based Question Answering**: As music is highly abstract, the model first analyzes fundamental attributes such as tonality, tempo, and emotion before progressing to genre classification and deeper inferential reasoning.

3.3 CoTA Dataset Analysis

A comprehensive statistical overview is presented in Table 1. To evaluate the quality and reasoning efficacy of the CoTA dataset, we analyze its design from two key perspectives: (1) **comprehensive audio coverage**, ensuring broad representation across real-world and synthetic scenarios, and (2) **scalability of reasoning complexity**, which aligns task difficulty with structured inference patterns.

Comprehensive Audio Coverage. CoTA spans speech (38.33%), music (14.12%), and environmental sounds (47.55%)—to ensure broad coverage of real-world auditory scenarios. It includes tasks from conversational speech (*e.g.*, speech-to-text translation tasks in *CoVoST 2*) to intricate musical structures (*MusicBench*) and fine-grained environmental sound analysis (*e.g.*, *AudioSet*'s rich descriptions of acoustic environments).

CoTA's key is its mix of real and synthetic data: 14.15% synthetic samples (*e.g.*, Multi-Speaker, Complex Audio) to support complex reasoning tasks, and most high-quality real-world datasets (*e.g.*, *MELD* for emotion recognition). By unifying 10 task types—from basic classification to advanced reasoning like translation and irony detection—CoTA enables hierarchical learning.

Scalability of Reasoning Complexity. The word count distribution reflects the capability of handing reasoning complexity. In Figure 6 of Sec. C, most responses range from 300 to 500 words, enabling detailed reasoning in audio and music QA. More complex tasks, like Multi-Speaker, can extend up to 1,500 words, showing the model's capacity to break down intricate, multi-element problems.

Simpler tasks such as S2TT require 100–200 words, ensuring no over-explaining. This adaptive response length highlights the model's flexibility in balancing depth and efficiency across tasks. More analysis about scalability is shown in Sec. C.

Table 1: CoTA composition. We adopt Google Gemini (Team et al., 2024) to build the reasoning ability in CoTA. Multi-Speaker and Complex Audio datasets are manually synthesized, details of which is in Sec. B.

Category	Dataset Source	Main Skills Learning	Model Used	Quantity	Percentage	Synthetic
Speech	Multi-Speaker	Multi-speaker Speech QA	gemini-2.0-flash	117.4k	12.09%	Yes
	MELD (Poria et al., 2019)	Speech Emotion QA	gemini-2.0-pro-exp	29.2k	3.01%	No
	CoVoST2 (Wang et al., 2021)	Speech-to-Text Translation	gemini-2.0-flash	224.6k	23.13%	No
Music	MusicBench (Melechovsky et al., 2024)	Music QA	gemini-2.0-flash	137.1k	14.12%	No
Sound	AudioSet (Gemmeke et al., 2017)	Sound QA	gemini-2.0-flash	315.2k	32.46%	No
	Clotho (Drossos et al., 2020)	Sound QA	gemini-2.0-pro-exp	9.3k	0.93%	No
	AudioCaps (Kim et al., 2019)	Sound QA	gemini-2.0-flash	117.5k	12.10%	No
	Complex Audio	Complex Audio QA	gemini-2.0-flash	20k	2.06%	Yes

Table 2: Evaluation summary for Audio-Reasoner.

Dataset	Split	Metric
MMAU-mini	Sound	ACC
	Speech	ACC
	Music	ACC
AIR-Bench	Chat-Sound	GPT-4 Eval
	Chat-Speech	GPT-4 Eval
	Chat-Music	GPT-4 Eval
	Chat-MixedAudio	GPT-4 Eval
	Foundation-SoundAQA	ACC
	Foundation-SER	ACC
	Foundation-SIC	ACC
	Foundation-SNV	ACC
	Foundation-MusicAQA	ACC
CoVoST 2	Test	BLEU
MELD	Test	ACC

4 Experimentation

4.1 Experimental Setup

Training Details. Our Audio-Reasoner, is based on Qwen2-Audio-Instruct (Chu et al., 2024) with 8.4 billion parameters. It was trained using the ms-swift framework (Zhao et al., 2024b) with full-parameter supervised fine-tuning, a peak learning rate of 1e-5, and a single epoch over the full CoTA.

Evaluation Metric. We first measure accuracy on closed-form questions using the MMAU-mini subset (Sakshi et al., 2024), because the model was not trained on multiple-choice data. We then assess real-world conversational ability via the chat and foundation sections of AIR-Bench (Yang et al., 2024b), covering sound, speech, and music modalities. We also evaluate speech-to-text translation on CoVoST 2 (Wang et al., 2021) and speech emotion recognition on MELD (Poria et al., 2019). Finally, we also test on the latest high-difficulty dataset MMAR (Ma et al., 2025b), one of the most challenging benchmarks in audio understanding and reasoning. Table 2 summarizes all evaluation datasets.

Baselines. We primarily select state-of-the-art large audio language models as the baselines for comparison. These include the closed-source models Gemini-1.5-pro (Team et al., 2024), GPT-4o (Hurst et al., 2024), Qwen-audio-turbo (Chu et al., 2023), as well as the open-source models SALMONN (Tang et al., 2023), Qwen-Audio-Chat (Chu et al., 2023), and Qwen2-Audio-Instruct (Chu et al., 2024) that also serves as the base model. Additionally, we compared cascade model approaches such as Whisper (Radford et al., 2023) + GPT-4 (Achiam et al., 2023) and a series of mainstream multimodal large language models (Gong et al., 2023b,a; Kong et al., 2024; Ghosh et al., 2024; Liu et al., 2024b; Su et al., 2023; Wu et al., 2024; Wang et al., 2023; Zhang et al., 2023b).

4.2 Main Results

Performance on MMAU-mini. Table 3 evaluates multimodal audio understanding on MMAU-mini across sound, music, and speech. Against closed-source models, Audio-Reasoner obtains the highest accuracy (61.71%), surpassing GPT-4o (57.30%) and Gemini-1.5-Pro (54.90%). The largest gain is in music reasoning (64.30% *vs* 60.77% and 49.40%). Speech reasoning is also strong (60.70% *vs* 53.15% and 58.55%). Compared to open-source models, Audio-Reasoner gains 12.51% over Qwen2-Audio-Instruct. It obtains 60.06% on sound (*vs*. 54.95%), 64.30% on music (*vs*. 50.98%), and 60.70% on speech (*vs*. 42.04%).

Performance on AIR-Bench chat. *(1) chat benchmark*. Table 4 evaluates contextual and conversa-

Table 3: Performance comparison on MMAU-mini. The **{so, mu, sp}** indicates whether "sound", "music", and "speech" have been used in training.

Model	Size	{so, mu, sp}	Sound	Music	Speech	Avg
Closed-Source						
gpt4o + caption	-	—	63.36	60.77	53.15	57.30
gemini-1.5-pro	-	—	56.75	49.40	58.55	54.90
Open-Source						
LTU	7B	Y Y N	22.52	9.69	17.71	16.89
LTU-AS	7B	Y Y Y	23.35	9.10	20.60	17.68
Audio Flamingo-Chat	2.2B	Y Y N	23.42	15.26	11.41	16.69
GAMA	7B	Y Y N	41.44	32.33	18.91	30.90
GAMA-IT	7B	Y Y N	43.24	28.44	18.91	30.20
MU-LLaMA	7B	N Y N	40.84	32.63	22.22	31.90
SALMONN	13B	Y Y Y	41.00	34.80	25.50	33.70
Qwen-audio-Chat	8.4B	Y Y Y	55.25	44.00	30.03	43.10
Qwen2-Audio-Instruct	8.4B	Y Y Y	54.95	50.98	42.04	49.20
Ours						
Audio-Reasoner	8.4B	Y Y Y	**60.06**	**64.30**	**60.70**	**61.71**

Table 4: Performance comparison on AIR-Bench Chat and Foundation benchmarks.

Model	Airbench-Chat					Airbench-Foundation					
	Sound	Music	Speech	Mixed Audio	Avg	Sound	Music	Sp-SER	Sp-SIC	Sp-SNV	Avg
Closed-Source											
Whisper+GPT4	-	-	7.54	-	7.54	-	-	59.5	87.7	30.0	59.1
Qwen-Audio-Turbo	6.59	5.98	7.04	5.77	6.34	62.8	62.5	60.0	56.4	54.3	59.2
Open-Source											
NEXT-GPT	4.76	4.18	3.86	2.92	4.13	18.8	47.1	25.7	25.6	25.4	28.5
SpeechGPT	0.95	0.95	1.57	1.14	1.15	33.9	31.3	37.6	45.8	32.6	36.2
BLSP	5.55	5.08	6.17	4.52	5.33	36.1	31.0	27.4	46.6	28.1	33.8
PandaGPT	5.46	5.06	3.58	2.93	4.25	48.7	50.7	26.0	28.5	43.2	39.4
SALMONN	6.28	5.95	6.16	6.08	6.11	28.4	54.6	29.9	36.7	34.3	36.8
Qwen-Audio-Chat	-	-	-	-	-	64.6	48.2	43.2	77.8	35.3	53.8
Ours											
Audio-Reasoner	**7.68**	**8.05**	**8.19**	**6.65**	**7.94**	**65.7**	**55.2**	**60.5**	**88.1**	**56.3**	**65.2**

Table 5: Performance comparison of the speech-to-text translation (S2TT) task on CoVoST 2 dataset. We consider the mutual conversion between Chinese and English as training and evaluation data.

Model	EN-ZN					ZN-EN					Avg
	BLEU1	BLEU2	BLEU3	BLEU4	Avg	BLEU1	BLEU2	BLEU3	BLEU4	Avg	
Closed-Source											
Gemini-1.5-pro	68.25	49.12	37.81	29.79	46.24	51.83	26.61	16.27	10.88	26.39	36.32
Open-Source											
Qwen2-Audio-Instruct	58.63	39.55	28.71	21.40	37.07	48.52	24.31	14.65	9.24	24.18	30.63
Ours											
Audio-Reasoner	**72.89**	**54.17**	**42.46**	**33.95**	**50.87**	**56.50**	**29.99**	**18.37**	**11.62**	**29.13**	**40.00**

tional reasoning on AIR-Bench chat across sound, music, speech, and mixed audio. Among closed-source models, Audio-Reasoner achieves the highest overall score (7.94), surpassing Gemini-1.5-Pro (5.70) and Whisper+GPT-4 (7.54), with notable gains in music (8.05) and speech (8.19). Its mixed audio score (6.65) also highlights strong multi-source audio reasoning. Compared to open-source models, Audio-Reasoner outperforms Qwen2-Audio by 1.01 points (7.94 *vs.* 6.93). It scores 7.68 on sound (*vs.* 6.99), 8.05 on music (*vs.* 6.79), and 8.19 on speech (*vs.* 7.18).

(2) foundation benchmark. Table 4 presents AIR-Bench foundation results. Audio-Reasoner achieves the highest overall score (65.2), outperforming both closed- and open-source baselines. It leads the best closed-source model, Qwen-Audio-Turbo (59.2), by 6.0 points, reflecting consistent strength across all domains. In sound reasoning, it scores 65.7, surpassing Qwen-Audio-Chat (64.6) and Qwen-Audio-Turbo (62.8), indicating robust

Table 6: Performance comparison of the speech emotion recognition (SER) task on MELD dataset.

Model	Unweighted_ACC
EMO-box	31.5
SALMONN	39.2
Qwen2-Audio-Instruct	49.9
Audio-Reasoner	53.9

non-speech audio understanding. For music, it achieves 55.2, outperforming Qwen-Audio-Turbo (48.2) and all open-source models, showcasing superior grasp of musical structure. In speech, Audio-Reasoner sets new records: 60.5 in SER (*vs*. 60.0), 88.1 in SIC (*vs*. 87.7), and 56.3 in SNV (*vs*. 54.3), with its SIC performance showing strong speaker recognition enabled by CoTA's stepwise reasoning.

Performance on CoVoST 2. Table 5 evaluates speech-to-text translation on CoVoST 2. Audio-Reasoner excels in both EN-ZN and ZN-EN translation. For *EN-ZN*, it achieves an average BLEU of 50.87, outperforming Gemini-1.5-Pro (46.24) by 4.63 points and Qwen2-Audio-Instruct (37.07) by 13.80. Its BLEU-4 of 33.95 reflects high fluency and quality in complex translations. In *ZN-EN*, it scores 29.13, surpassing Gemini-1.5-Pro (26.39) and Qwen2-Audio-Instruct (24.18). A BLEU-4 of 11.62 further highlights its coherence in challenging, long-form outputs. These results confirm Audio-Reasoner's superior cross-lingual alignment and translation performance.

Performance on MELD. Table 6 reports results on MELD for speech emotion recognition. Audio-Reasoner achieves the highest unweighted accuracy (53.9), surpassing the previous best, Qwen2-Audio-Instruct (49.9), by 4.0 points. This shows its strong emotion perception and contextual understanding.

Why Audio-Reasoner Excels. A key factor driving Audio-Reasoner's performance is the CoTA dataset, which provides comprehensive, structured, and context-aware audio reasoning. *(1) CoTA's comprehensive audio reasoning data*. Unlike models trained on fragmented or task-specific datasets, Audio-Reasoner benefits from CoTA's extensive and balanced coverage of sound, speech, and music-based reasoning. *(2) CoTA's strong logical and contextual inference*. CoTA's emphasis on structured reasoning and contextual awareness enables Audio-Reasoner to outperform existing models in long-form reasoning (MMAU-mini) and conversational audio understanding (AIR-Bench

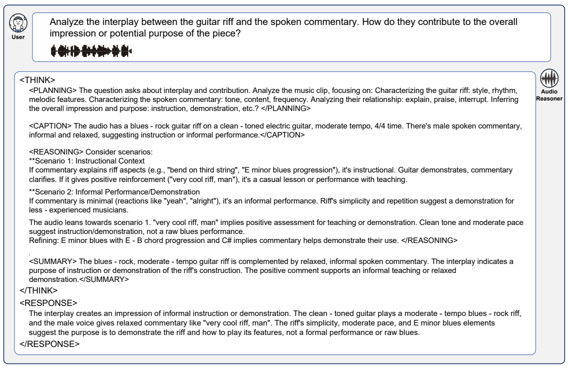

Figure 4: Example of Audio-Reasoner answering music-based question. Best view with zooming in.

chat). These strengths position Audio-Reasoner as a breakthrough in open-source audio intelligence. In the latest audio-based reasoning benchmark MMAR(Ma et al., 2025b) , Audio-Reasoner still achieves the best results among all large audio-language models, which also demonstrates the advantages of CoTA. We have included the relevant results in Sec. D.

4.3 Case Study

Here we show a case in Figure 4, demonstrating the audio-based reasoning capability of Audio-Reasoner. The system analyzes the interplay between a guitar riff and a spoken commentary, systematically breaking down their characteristics and relationship. It identifies key musical features, assesses the commentary's tone and intent, and infers the overall purpose of the piece. By considering different scenarios, Audio-Reasoner determines that the interplay suggests an informal instructional or demonstrative context rather than a formal performance. This example highlights the model's ability to extract meaningful insights from audio, combining musical analysis with contextual interpretation. See Sec. E for more examples.

4.4 Error Analysis

To better understand the limitations of Audio-Reasoner, we conducted a comprehensive error analysis on a subset of incorrectly answered questions from the MMAU-mini benchmark. Table 7 presents representative examples of different error categories, illustrating both the erroneous reasoning paths taken by our model and the correct reasoning paths.

As summarized in Table 8, the dominant failure mode arises from perceptual errors (49%), indicating that nearly half of all mistakes emerge when the

Table 7: Representative errors and reasoning paths across different failure modes in Audio-Reasoner.

Error Type	Erroneous Reasoning Path	Correct Reasoning Path
Perceptual Error (mis-hearing & acoustic segmentation failure)	**Prompt**: "Audio clip: estimate the tempo (BPM) of the drumming pattern." Model hears an 8-bar loop as 4 bars → computes 4 bars / 10 s = 0.4 bars/s → BPM≈96 → chooses B (90–100 BPM)	The clip actually contains 8 bars in 10 s (with ghost notes & reverb) → 8 bars/10 s = 0.8 bars/s → BPM≈192 → choose D (190–200 BPM)
Knowledge Error (missing domain facts or unit conventions)	**Prompt**: "A bat emits an ultrasonic chirp at 40 kHz; echo returns after 6 ms. How far away is the insect?" Model uses distance = speed × time, neglects round-trip → 343 m/s × 0.006 s ≈ 2.06 m → picks C (≈2 m)	Account for round-trip: distance = $\frac{1}{2} \cdot v \cdot t = 0.5 \times 343 \times 0.006 \approx$ 1.03 m → pick B (≈1 m)
Reasoning Error (algebraic or arithmetic slip)	**Prompt**: "Two successive echoes from a canyon wall arrive at 1.2 s & 3.6 s; what is the wall distance?" Model sets $2d/v = 3.6 - 1.2$ correctly but mis-types $v = 300$ as 30 → $d \approx$ 36 m → picks A (0–50 m)	Using $v \approx$ 343 m/s: $d = v \cdot (\Delta t)/2 \approx 343 \times 2.4/2 \approx 411$ m → pick D (400–450 m)
Other (Instruction Mis-read) (format or mapping mistake)	**Prompt**: "Count distinct speakers in the dialogue." True answer: 4 speakers. Model correctly separates voices but writes "There are four speakers" → mistakenly selects choice "3" (off-by-one in mapping)	Maps verbal answer to the correct choice label "4"

Table 8: Error distribution analysis on MMAU-mini benchmark failures.

Error Type	Percentage
Perceptual Errors	49%
Knowledge Errors	40%
Reasoning Errors	3%
Wrong/Misunderstanding Instruction	3%
Choice Format Error	2%
Repeat Thinking and No Answer	3%

model cannot execute core auditory-understanding skills such as discriminating sounds, parsing musical timbre, or transcribing speech. These lapses expose the system's limited ability to track fine-grained acoustic cues and maintain robust representations under noisy or multi-speaker conditions. A further 40% of the errors belong to the knowledge category, underscoring the challenge of applying domain-specific facts and conventions in audio reasoning tasks. Together, these two categories account for almost nine-tenths of all failures, revealing a bottleneck in aligning perception with knowledge application.

5 Conclusion

In this work, we introduced Audio-Reasoner, a large audio language model (LALM) designed to advance deep reasoning in audio-based tasks. By leveraging inference scaling and structured chain-of-thought (CoT) reasoning, we demonstrated significant performance improvements across key benchmarks. Central to our approach is CoTA, a large-scale, high-quality dataset containing around 1.2 million structured reasoning samples, which we generated through a systematic pipeline of annotation refinement, question synthesis, and CoT generation. Our results show the efficacy of structured reasoning in the audio domain, achieving state-of-the-art performance on MMAU-mini (+25.04%), CoVoST 2 (+8.31%), and MELD (+8.01%). These findings underscore the critical role of reasoning-rich datasets and inference scaling in multimodal learning, particularly for audio-based tasks where existing models struggle with complex reasoning.

6 Limitations

While Audio-Reasoner shows promising performance, limitations remain. First, the current model primarily handles single-turn reasoning and may struggle with more complex multi-turn or contextual scenarios, where maintaining context over time is crucial. Second, its generalization ability across diverse audio domains and real-world noise conditions requires further validation, as its performance in noisy or varied environments has yet to be fully tested. Additionally, cross-modal reasoning—especially integrating visual or textual cues—remains unexplored, limiting its potential in multi-modal applications. Future work will focus on addressing these limitations.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games (T-CIAIG)*, (1):1–43.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Yingqian Cui, Pengfei He, Xianfeng Tang, Qi He, Chen Luo, Jiliang Tang, and Yue Xing. 2024. A theoretical understanding of chain-of-thought: Coherent reasoning and error-aware demonstration. *arXiv preprint arXiv:2410.16540*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and Wynne Hsu. 2024. Video-of-thought: step-by-step video reasoning from perception to cognition. In *International Conference on Machine Learning (ICML)*, pages 13109–13125.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Empirical Methods in Natural Language Processing*, pages 6288–6313.

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023a. Joint audio and speech understanding. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023b. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

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

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. 2024. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. 2024. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. *arXiv preprint arXiv:2411.14794*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Feihu Jin, Yifan Liu, and Ying Tan. 2024. Zero-shot chain-of-thought reasoning guided by evolutionary algorithms in large language models. *arXiv preprint arXiv:2402.05376*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 119–132.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *International Conference on Machine Learning (ICML)*, pages 25125–25148.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a.

Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024b. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290.

Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. 2024. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. *arXiv preprint arXiv:2406.07162*.

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

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

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 8286–8309.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 527–536.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pages 28492–28518.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 53728–53741.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations (ICLR)*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. Chain of thoughtlessness? an analysis of cot in planning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 29106–29141.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. In *Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants (TLLM)*, pages 11–23.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Yunlong Tang, Gen Zhan, Li Yang, Yiting Liao, and Chenliang Xu. 2024. Cardiff: Video salient object ranking chain of thought reasoning for saliency prediction with diffusion. *arXiv preprint arXiv:2408.12009*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 74952–74965.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *Conference of the International Speech*

Communication Association (Interspeech), pages 2247–2251.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2023. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2024. Videocot: A video chain-of-thought dataset with active annotation tool. In *Workshop on Advances in Language and Vision Research (ALVR)*, pages 92–101.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 24824–24837.

Peng Wen, Teng-Gen Hu, Robert J Linhardt, Sen-Tai Liao, Hong Wu, and Yu-Xiao Zou. 2019. Mulberry: A review of bioactive compounds and advanced processing technology. *Trends in food science & technology*, 83:138–158.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *International Conference on Machine Learning (ICML)*, pages 53366–53397.

Zhifei Xie and Changqiao Wu. 2024a. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Zhifei Xie and Changqiao Wu. 2024b. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024a. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024b. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1979–1998.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11809–11822.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 15757–15773.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023b. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024a. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. 2024b. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024a. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. 2024b. Swift: a scalable lightweight infrastructure for fine-tuning. *arXiv preprint arXiv:2408.05517*.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*.

Anni Zou, Zhuosheng Zhang, Hai Zhao, and Xiangru Tang. 2023. Generalizable chain-of-thought prompting in mixed-task scenarios with large language models. *arXiv preprint arXiv:2310.06692*.

A Prompt Details

A universally applicable method for writing prompts involves three key components: a clear task definition, a structured example, and a precise format specification. Our prompt adheres to this methodology by first defining the task explicitly, outlining the need for detailed audio descriptions and progressively challenging questions. It then provides a structured example that demonstrates the expected output format, ensuring clarity and minimizing ambiguity. Lastly, it specifies the exact formatting rules using delimiters such as <caption>...</caption> and <question1>...</question1>, ensuring consistency in responses. This approach guarantees efficiency by eliminating interpretative variance, allowing for precise and reproducible outputs. When drafting this prompt, we adhered to a structured approach to maximize clarity and effectiveness. The first-person perspective is used to emphasize our direct involvement in designing the task, ensuring the reader understands the rationale behind each structural choice. The structure follows a logical progression: we begin by introducing the general method, transition into an explanation of how our prompt aligns with this method, and conclude by justifying the approach's efficiency. By maintaining an academic tone, we reinforce the credibility and rigor of our prompt-writing methodology.

A.1 Prompt of Stage 1 when Processing Data (Sample from AudioSet)

We are annotating some audio and designing some questions. You are an excellent audio analyst. Next, you will receive an audio and one absolutely correct but simple description. Your task is to first generate a more detailed, in-depth and absolutely correct new description based on the given descriptions. Then, use this description to generate three open-ended or single-choice questions with four options along with their answers. Please separate different parts using <caption>...</caption> <question1><question>...</question> <answer>...</answer></question1> <question2> <question>...</question> <answer>...</answer></question2>......

Here is a sample. Please strictly follow the format in the sample. <caption>The audio presents a sustained, high-frequency static noise, characteristic of a detuned or malfunctioning electronic device, likely a television or radio...</caption><question1><question>Describe the characteristics of the static noise in the audio, and how these characteristics change over time.</question><answer>...</answer></question1><question2> <question>What...?</question> <answer>...</answer> </question2> <question3><question>What...?</question> <answer>...</answer> </question3>

Here is the original description: ***** label here *****.
Here is the audio.
Please design three questions that gradually become more challenging, starting from basic factual questions, but don't deviate from the content of the audio itself. If it's a single-choice question, please give four options like (a) one, (b) two, ... and the answer should be analyzed and end with a format like "the answer is B."

A.2 Prompt of Stage 2 when Processing Data (Sample from AudioSet)

We are now designing a system to generate structured audio-based chain-of-thought reasoning data. You will receive an audio clip, its textual description, as well as a question and its answer. Your task is to explore in more detail the thinking process from the question to the answer. Your chain of thought will be divided into two main parts, the <think> and <response> sections. In the <think> section, you need to go through four steps: planning, captioning, reasoning, and summarizing. The <think> section is invisible to the user. Therefore, in the <response> section, you need to base on all the reasoning processes and results in the <think> section and provide a final reply based on the question. Finally, your reply should strictly follow the following format: <THINK><PLANNING> (In this part, analyze the user's needs and how to complete the task. If the problem is complex, it should be done step by step) </PLANNING><CAPTION> (In this part, conduct an overall analysis of the given audio input content, try to find all the parts related to the question, describe them, and ensure it is completely correct.) </CAPTION><REASONING> (Start reasoning towards the final answer here) </REASONING><SUMMARY> (Draw appropriate results based on the reasoning part here) </SUMMARY></THINK><RESPONSE> Give the final answer here referring to the <THINK> part </RESPONSE> Please strictly follow the format of the sample. Sample:
<THINK>
<PLANNING>
The user wants to understand the dynamic changes within the provided audio clip ......
</PLANNING>
<CAPTION>
The audio clip predominantly features static noise. ...... similar to that of a detuned television or a device failing to receive a signal.
</CAPTION>
<REASONING>
1. Identify changes in Intensity (Volume): The audio's static noise does not remain at a constant volume. There are noticeable increases and decreases in loudness throughout the clip. ...... is struggling to maintain a consistent output, adding to the impression of something malfunctioning or broken.
</REASONING>
<SUMMARY>
The static noise in the audio is highly dynamic. ...... leading to a sense of disorder and instability.
</SUMMARY>
</THINK>
<RESPONSE>
The audio presents a static noise, ...The overall effect is one of energetic chaos, preventing any possibility of calm or predictability.
</RESPONSE>
Note that you have both the question and the answer because it is necessary to ensure the correctness of the chain of thought. However, in your response, you can only refer to the content of the question and the audio, and lead to the answer. You must absolutely not assume that you already know the answer. Please provide a detailed and flexible response with high-quality logic in both the caption and reasoning sections. If the reasoning part requires complex logic, you can even propose several different approaches and try them one by one.
Here is the original description: ***** caption here *****.
The question is: ***** question here *****.
The answer you can refer to : ***** answer here *****.
Again, don't mention that you have the answer and the description because they are only here to help you to design the chain of thought but should not exist in the real-world scenario, either in the think or response sections.

A.3 Prompt of Stage 3 when Processing Data (Sample from AudioSet)

We are data reviewers. Next, you will receive an audio clip, along with its description, questions, answers, and most importantly, the thought process for solving the problems. Please determine and analyze whether all of these elements are completely correct, especially check if there are any hallucinations in the thought process. Return <True> if there are no issues, and <False> if there are errors in the data.
Here is the description of the audio: *** **caption here** ***.
Here is the question: *** **question here** ***.
Here is the answer: *** **answer here** ***.
And here is the thought process: *** **COT process here** ***.
Please conduct a thorough judgment and analysis and provide the result in the specified format.

B Synthetic Data Generation Pipeline

B.1 Synthetic Data Introduction

Multi-Speaker Dataset: To enhance the model's ability to comprehend complex, multi-turn conversations among multiple speakers, we constructed the Multi-Speaker dataset using text-to-speech (TTS) technology. The dataset generation process consists of three steps: (1) **Conversation Text Generation**: We utilized commericial models to generate diverse multi-speaker conversation texts covering a wide range of scenarios. (2) **Speech Synthesis**: Leveraging all available timbres from LibriSpeech (Panayotov et al., 2015) as prompts, we employed the CosyVoice2 (Du et al., 2024) framework to synthesize high-quality speech samples. (3) **Dataset Assembly**: The synthesized speech samples, feauring distinct timbres, were carefully combined to create a rich and diverse multi-speaker conversation dataset.

This approach ensures that the dataset effectively trains the model to navigate intricate speaker interactions in natural conversations.

Complex Audio Dataset: To strengthen the model's ability to process intricate audio compositions, we developed the Complex Audio dataset. The audio materials are sourced from AudioSet, AudioCaps, and our internal material library. The dataset consists of two primary types of audio data: (1) **Temporally Complex Audio**: Short audio clips were sequentially arranged, requiring the model to perform tasks such as sorting or counting based on temporal relationships. We sequentially arranged different short audio clips and required the model to perform sorting or counting tasks. (2) **Composite Complex Audio**: Multiple long-form audio tracks—such as engine sounds and musical instruments–were blended. The model was then tasked with extracting specific target sounds from the composite audio, enhancing its ability to analyze overlapping audio sources.

This dataset ensures that the model is well-equipped to handle real-world audio challenges, such as distinguishing between multiple concurrent sounds and identifying temporal patterns.

To better understand what Multi-Speaker dataset and Complex Audio dataset are respectively tasked with, we provide some examples in Sec. B.2, Sec. B.3.

B.2 Sample from Complex Audio Dataset

<THINK>
<PLANNING>
The question is about identifying the sounds in the audio and arranging them in the correct order. I will carefully listen to the audio multiple times, noting down the sounds as I recognize them. Then, I will compare the identified sound sequence with the provided answer options to choose the one that matches best.
</PLANNING>
<CAPTION>Upon initial listening, I can discern sounds that seem like a shattering noise, an animal sound which I think sounds like a cow, a cat sound, and what sounds like some mechanical tools working, perhaps a chainsaw. I shall listen again to clarify the sounds and their order.
</CAPTION>
<REASONING>1. After relistening, the first sound is definitely glass breaking.
2. The second sound is definitely a cow mooing.
3. The third sound is a cat meowing.
4. The fourth sound is a chainsaw.
5. I compare this sequence (glass break, cow, cat, chainsaw) with the options provided.
6. Option (a) matches the order I identified.
</REASONING>
<SUMMARY>Based on the detailed listening and analysis, the correct order of sounds in the audio is: glass break, cow, cat, chainsaw. This matches option (a).
</SUMMARY>
</THINK>
<RESPONSE>
The answer is a.
</RESPONSE>

B.3 Sample from Multi-Speaker Dataset

<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>

C Further Dataset Analysis

C.1 More Statistical Results of CoTA

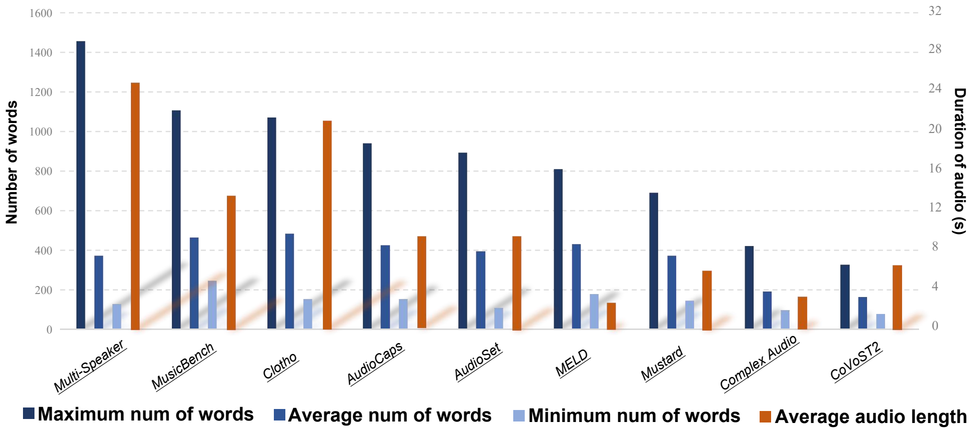

Figure 5: Detailed information of the CoTA dataset: The maximum, minimum, and average number of words in the answers, as well as the average length of the audio.

CoTA's reasoning adaptability is also evident in its distribution of token lengths and audio duration across tasks in Figure 5 . The dataset accommodates a broad range of reasoning complexities, with token lengths spanning 327 (CoVoST2) to 1,454 (Multi-Speaker), ensuring coverage of both concise and highly intricate reasoning processes. Notably, tasks requiring deep logical inference, such as complex audio, exhibit a well-balanced token distribution (max = 423, avg = 192.96), supporting structured multi-step reasoning without unnecessary redundancy.

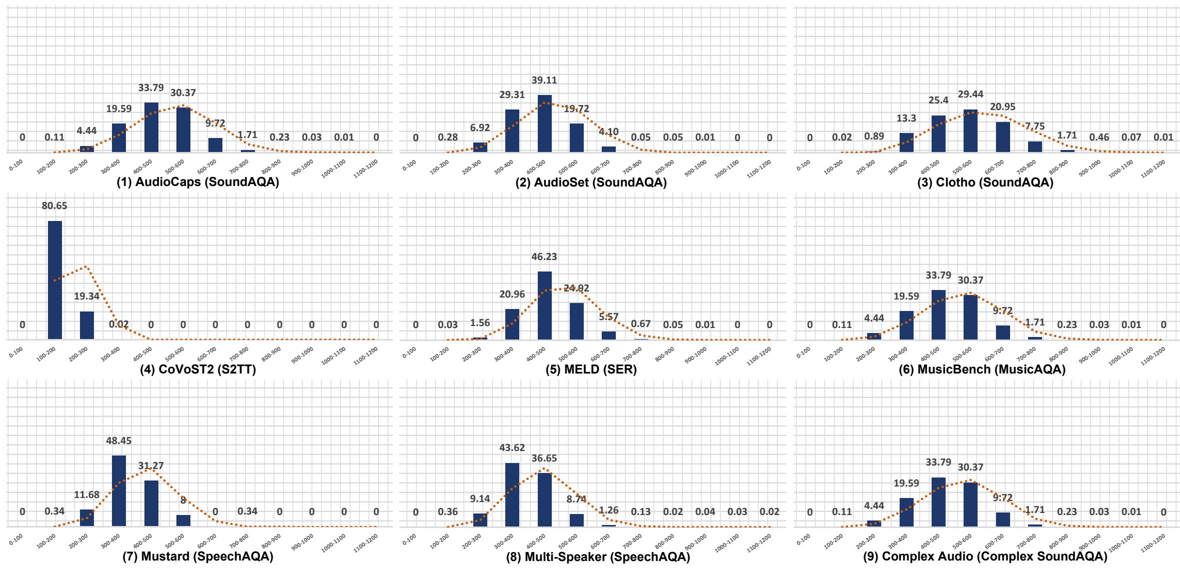

Figure 6: The bar chart shows the data length distribution across nine CoTA sub-datasets, with intervals of 100 on the horizontal axis and proportions on the vertical axis. A moving average trend line is overlaid.

Further, the dataset's average token lengths (164.48–481.57) align with task difficulty: longer reasoning chains characterize tasks such as sound description (AudioSet: 395.26) and music understanding (MusicBench: 463.89). Meanwhile, CoTA ensures practical generalization by maintaining audio durations between 2.85s and 26.34s, where shorter clips (*e.g.*, MELD: 2.84s) support concise context-dependent reasoning, while extended sequences (Multi-Speaker: 26.34s) enable complex multi-turn inference. This systematic variation in reasoning depth and audio granularity ensures adaptability across diverse tasks, addressing the limitation of one-size-fits-all reasoning chains in existing audio datasets.

D Results on Latest Audio-based Reasoning Benchmark

Table 9: Performance comparison on MMAR.

Models	Size	Sound	Music	Speech	Sd-Mu	Sd-Sp	Mu-Sp	Sd-Mu-Sp	Avg (%)
Closed-Source									
GPT-4o mini Audio	-	**38.79**	**35.97**	**58.84**	**45.45**	**60.09**	**57.32**	**50.00**	**50.60**
Open-Source									
Audio Flamingo	2.2B	32.73	21.84	24.83	18.18	30.28	24.39	25.00	26.60
Audio Flamingo 2	3B	24.85	18.48	17.28	18.18	20.25	16.67	8.33	17.58
LTU	7B	19.39	19.00	19.35	19.18	24.77	21.95	16.67	19.20
LTU-AS	7B	20.00	18.09	9.09	9.09	20.64	20.85	12.50	15.18
GAMA	7B	29.09	24.27	27.89	27.27	24.77	28.05	20.83	26.50
GAMA-IT	7B	22.42	16.02	18.47	9.09	14.63	12.50	17.40	15.22
Qwen2-Audio	8.4B	28.18	22.60	32.35	9.09	26.34	28.63	33.33	25.79
Qwen2-Audio-Instruct	8.4B	33.33	24.27	32.31	9.09	31.19	30.49	40.91	28.95
SALAMONN	7B	30.91	31.07	**34.69**	9.09	30.85	27.08	37.50	30.43
SALAMONN	13B	30.30	31.07	**34.69**	9.09	34.86	**35.37**	41.67	33.20
Large Audio Reasoning Models (LARMs)									
Audio-CoT	8.4B	35.76	25.24	30.41	9.09	30.73	30.49	37.50	31.30
Ours									
Audio-Reasoner	8.4B	**43.64**	**33.50**	32.99	**45.45**	**42.66**	31.71	25.00	**36.80**

Performance on MMAR. When examining the average performance on MMAR(Ma et al., 2025b), Audio-Reasoner demonstrates remarkable superiority. It outperforms the highest-performing open-source model, SALAMONN 13B (with an average performance of 33.20%), by 3.60%, achieving an average of 36.80%. Looking at individual metrics, for the "Sound" category, Audio-Reasoner attains 43.64%, significantly surpassing Qwen2-Audio-Instruct which only reaches 33.33%. In the "Music" category, with a score of 33.50%, it outperforms SALAMONN 7B and 13B that both score 31.07%. In the "Speech" category, SALAMONN 7B and 13B score 34.69%. while Audio-Reasoner scores 32.99%, the overall performance still showcases Audio-Reasoner's strength. However, Compared with the closed-source model, GPT-4o mini Audio (with an average of 50.60%), there is a notable gap in Audio-Reasoner's results. This highlights that the open-source community still has substantial work ahead to enhance the reasoning capabilities of models.

E More Case Studies

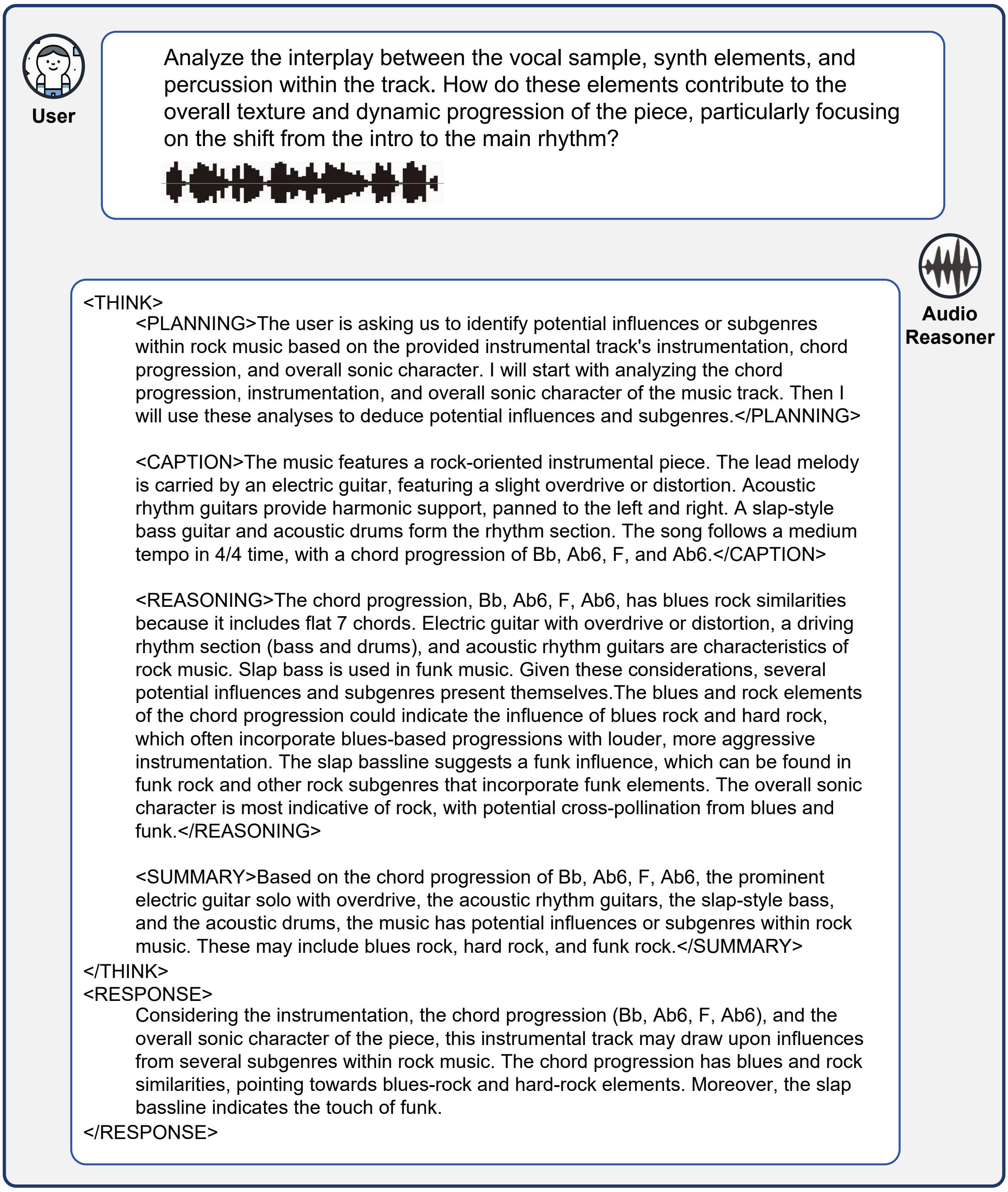

Figure 7: Example of Audio-Reasoner answering music-based question.

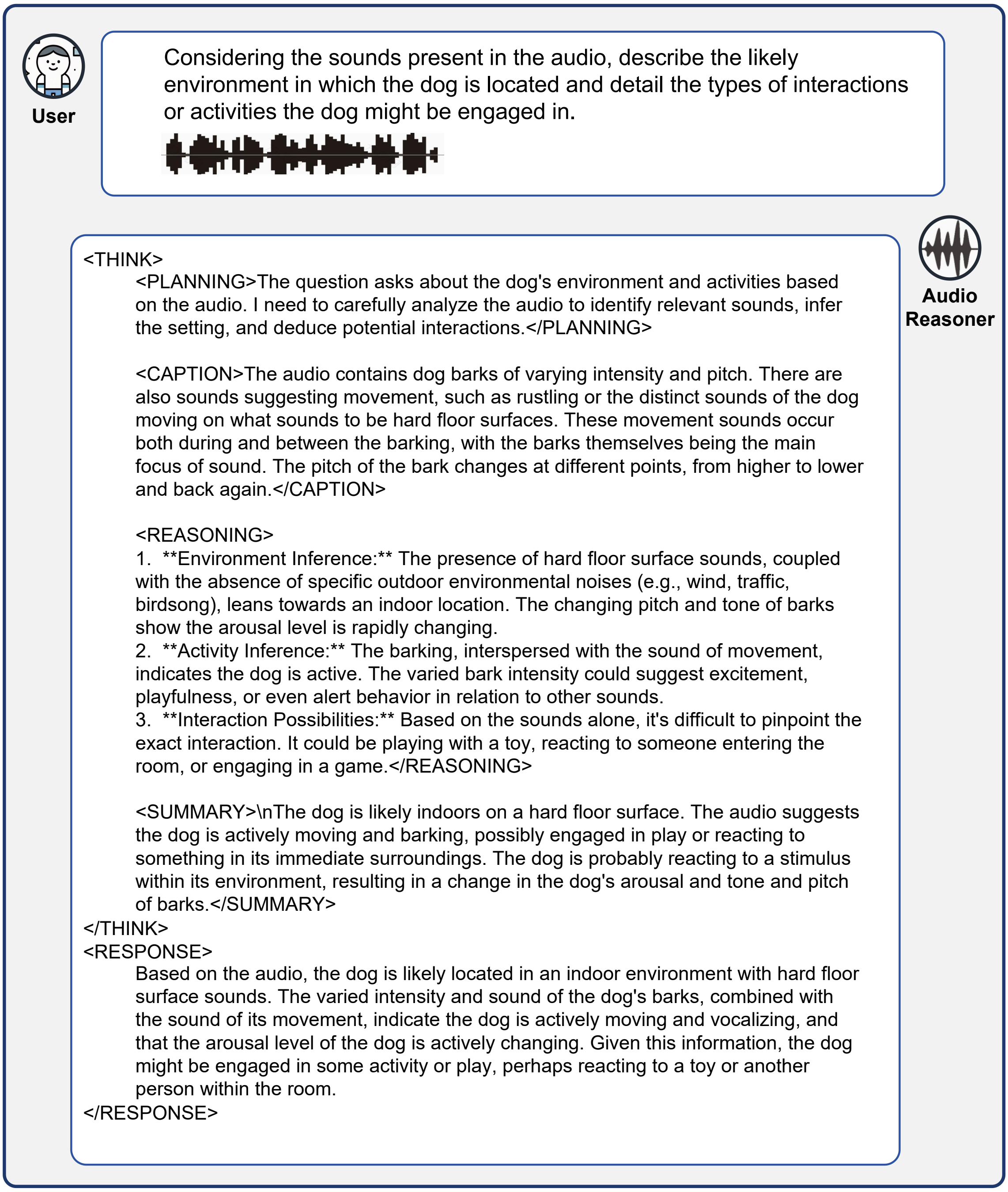

Figure 8: Example of Audio-Reasoner answering sound-based question.

User

What did the second person say about the rose bush?
(a) They like it.
(b) It's the bane of their existence.
(c) It needs more fertilizer.
(d) It provides good shade.

Audio Reasoner

<THINK>

<PLANNING>The question asks about what the second person said about the rose bush. The audio presents a conversation between two neighbors, Amelia and Ben. I need to identify who the second speaker is and what they said about the rose bush. Then, I need to match what the second speaker said with the given options.</PLANNING>

<CAPTION>The audio presents a conversation between Amelia and Ben about the upcoming block party and a troublesome rose bush.
Speaker 1 (Amelia): Asks Ben if he's attending the block party and mentions the rose bush causing him trouble. She says \"You mentioned weedies, that rose bush still causing you trouble, the one that keeps growing over the fence?\" She also mentions she has been scratched by it.
Speaker 2 (Ben): Confirms he's attending the block party, entering his chili in the cook-off, and complains about the rose bush. He says \"That thorny rose bush is the bane of my existence. It's constantly scratching me when I am gardening. So, yes, it's still causing me trouble.\" and suggests chopping it down.\n\nTherefore, Ben is the second speaker.</CAPTION>

<REASONING>
1. **Identify the second speaker:** Based on the audio, Ben is the second speaker.
2. **Identify what Ben said about the rose bush:** Ben says \"That thorny rose bush is the bane of my existence. It's constantly scratching me when I am gardening. So, yes, it's still causing me trouble.\" and suggests chopping it down.
3. **Match Ben's statement with the options:**
- (a) They like it. - This is incorrect because Ben clearly dislikes the rose bush.
- (b) It's the bane of their existence. - This aligns perfectly with Ben's statement.
- (c) It needs more fertilizer. - This is not mentioned in the audio.
- (d) It provides good shade. - This is also not mentioned in the audio.
Therefore, the correct answer is (b).</REASONING>

<SUMMARY>\nBased on the audio, the second speaker, Ben, describes the rose bush as \"the bane of my existence\" and complains about it scratching him. Therefore, the answer is (b). </SUMMARY>
</THINK>
<RESPONSE>
The second person, Ben, explicitly states that the rose bush \"is the bane of my existence\" . Therefore, the answer is B.
</RESPONSE>

Figure 9: Example of Audio-Reasoner answering multi-speaker speech-based question.

User

Given that you hear buzzing plastic, identify the background sound?
(a). A washing machine running
(b). Rain falling on a window
(c). A toilet flushing
(d). Waves crashing on a shore

Audio Reasoner

<THINK>
<PLANNING>
I will carefully and methodically approach the task of identifying the background sound in the audio clip. This will allow me to focus solely on the characteristics of the background noise. Once the background sound is isolated, I will analyze it in detail to determine its key qualities, such as its consistency, rhythm, and texture. Afterward, I will compare these characteristics to the provided answer options .</PLANNING>

<CAPTION> Upon my initial listening, I observed that the background sound possesses characteristics that suggest fluid movement. There is a continuous, almost rhythmic, sound with subtle gurgling and sloshing. This could indicate the presence of water or some other flowing liquid. As I continue to listen more carefully, the flow becomes even more evident, suggesting a consistent rush or stream-like sound. </CAPTION>

<REASONING>
1. **Sound Characteristics:** My first observation is that the background sound features a continuous, almost rhythmic quality, with faint gurgling noises. These characteristics often point to a fluid in motion. Upon further listening, I recognize that the sound might be created by a liquid flowing from one place to another.
2. **Detailed Listening:** I focus on the quieter sections of the sound, where the fluid dynamics become clearer. The consistent, almost rushing quality suggests the movement of water rather than irregular drops or splashes.
3. **Comparison to Options:** With this in mind, I now compare the sound characteristics to the provided answer options. - "A toilet flushing," which often involves a steady rush of water with gurgling and sloshing as the water flows through pipes. This closely matches the consistent flow and gurgling noises I hear in the background.
4. **Conclusion:** Based on my analysis, the sound best aligns with the description of a toilet flushing, due to the nature of the consistent, rushing water and subtle gurgling sounds. </REASONING>
</THINK>
<RESPONSE>
After listening to the audio carefully, isolating the background sound, and comparing it to the answer options, I concluded that the background sound most closely matches the sound of a toilet flushing. The key factors leading to this conclusion include the continuous rush of water and the gurgling quality, which are typical of a flushing toilet.
</RESPONSE>

Figure 10: Example of Audio-Reasoner answering complex-audio-based question.