SoundMind: RL-Incentivized Logic Reasoning for Audio-Language Models

Xingjian Diao[1], **Chunhui Zhang**[1], **Keyi Kong**[2], **Weiyi Wu**[1], **Chiyu Ma**[1],
Zhongyu Ouyang[1], **Peijun Qing**[1], **Soroush Vosoughi**[1], **Jiang Gui**[1]
[1]Dartmouth College, [2]Shandong University
xingjian.diao.gr@dartmouth.edu

Abstract

While large language models have shown reasoning capabilities, their application to the audio modality, particularly in large audio-language models (ALMs), remains significantly underdeveloped. Addressing this gap requires a systematic approach, involving a capable base model, high-quality reasoning-oriented audio data, and effective training algorithms. In this study, we present a comprehensive solution: we introduce the Audio Logical Reasoning (ALR) dataset, consisting of 6,446 text-audio annotated samples specifically designed for complex reasoning tasks. Building on this resource, we propose SoundMind, a rule-based reinforcement learning (RL) algorithm tailored to endow ALMs with deep bimodal reasoning abilities. By training Qwen2.5-Omni-7B on the ALR dataset using SoundMind, our approach achieves state-of-the-art performance in audio logical reasoning. This work highlights the impact of combining high-quality, reasoning-focused datasets with specialized RL techniques, advancing the frontier of auditory intelligence in language models. Our code and the proposed dataset are available at https://github.com/xid32/SoundMind.

1 Introduction

Large language models (LLMs) have made remarkable strides in reasoning capabilities through innovations such as Chain-of-Thought (CoT) prompting and specialized reasoning architectures. Recent models such as OpenAI's o1 model (Jaech et al., 2024) and Deepseek-R1 (Guo et al., 2025) have demonstrated exceptional performance on complex mathematical and programming tasks (Zhao et al., 2024; Team et al., 2025; Muennighoff et al., 2025). Particularly noteworthy is Deepseek-R1's rule-based reinforcement learning approach, which enables emergent reasoning without relying on traditional frameworks like Monte Carlo Tree Search (Wan et al., 2024) or process reward models (Lightman et al., 2023). This general reasoning paradigm has successfully extended to the visual domain, where frameworks like Visual-CoT (Shao et al., 2024) have significantly enhanced multimodal models' cognitive abilities for image and video reasoning.

Despite these advances in text and visual reasoning, a significant gap exists in *audio reasoning and generation capabilities*. While audio-language models (ALMs) like Audio Flamingo (Kong et al., 2024) and Qwen2.5-Omni-7B (Xu et al., 2025) have made progress in audio understanding, end-to-end audio reasoning remains underdeveloped. This limitation stems primarily from two factors: (1) the simplicity of existing audio reasoning datasets (Suzgun et al., 2023; Kong et al., 2024), which often contain only brief textual labels without proper audio modality annotations, and (2) the technical challenges in maintaining reasoning coherence during long-duration audio generation. Current CoT methods applied to audio often lead to hallucinations and performance degradation when generating extended reasoning sequences. The lack of aligned audio-text annotations further impedes research on reasoning-driven audio generation, creating a bottleneck in developing ALMs with sophisticated reasoning capabilities.

To address these challenges in audio reasoning (Figure 1), we introduce a comprehensive approach with two key components. First, we construct the Audio Logical Reasoning (ALR) dataset (Liu et al., 2023), containing 6,446 dual-modality samples with user content, CoT reasoning steps, final answers, and corresponding audio input/output. Derived from the Logi-QA 2.0-NLI dataset, ALR provides complete logical reasoning tasks with both content and response in text and audio modalities. Second, we propose the SoundMind algorithm, which builds upon the ALR dataset to address performance issues in long-form reasoning audio generation. Drawing inspiration from the Logic-RL

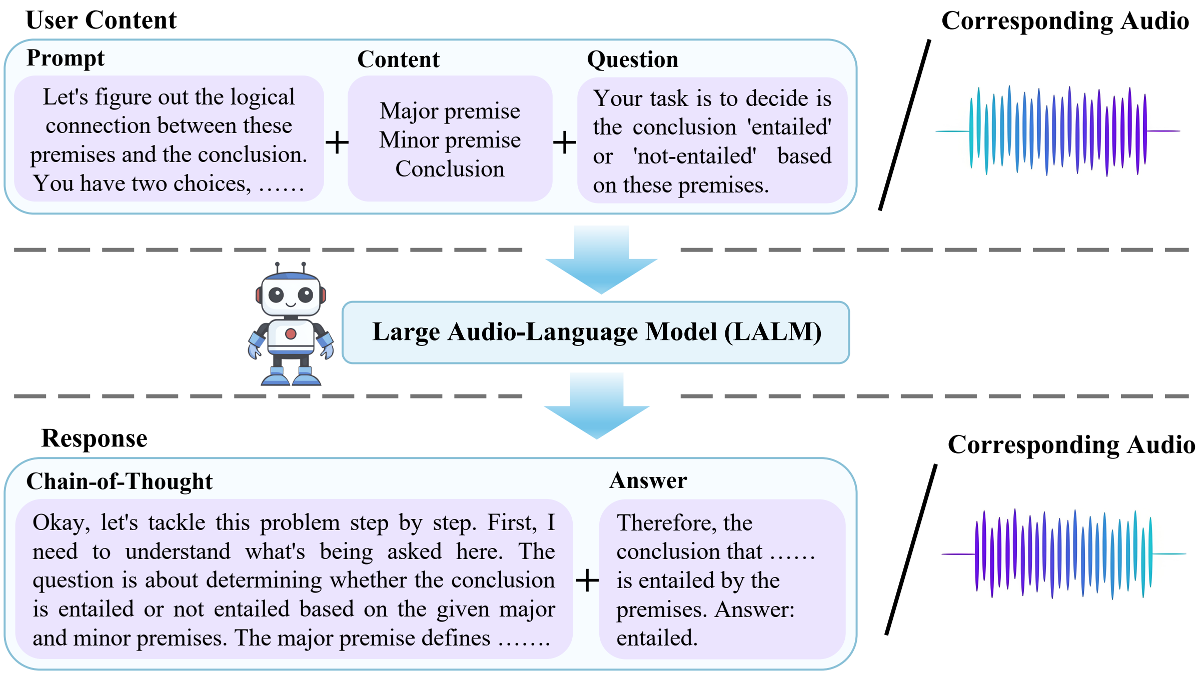

User Content
Prompt
Let's figure out the logical connection between these premises and the conclusion. You have two choices, ……
+
Content
Major premise
Minor premise
Conclusion
+
Question
Your task is to decide is the conclusion 'entailed' or 'not-entailed' based on these premises.
Corresponding Audio
Large Audio-Language Model (LALM)
Response
Chain-of-Thought
Okay, let's tackle this problem step by step. First, I need to understand what's being asked here. The question is about determining whether the conclusion is entailed or not entailed based on the given major and minor premises. The major premise defines …….
+
Answer
Therefore, the conclusion that …… is entailed by the premises. Answer: entailed.
Corresponding Audio

Figure 1: Overview of the end-to-end bimodal Natural Language Inference (NLI) task enabled by large audio-language models. The model receives input and produces output in either pure text or pure audio format.

framework (Xie et al., 2025a), SoundMind incorporates a strict format-based reward mechanism to prevent shortcut reasoning biased toward the textual modality. It leverages the REINFORCE++ algorithm (Hu, 2025) alongside the reward design principles from Deepseek-R1 for effective post-training. Our main contributions are as follows:

- **ALR Dataset:** We release a high-quality dual-modality dataset containing 6,446 samples with user content, CoT reasoning, answers, and corresponding audio input/output. The ALR dataset includes deep reasoning annotations in both text and audio formats, providing a valuable resource for RL-based training of audio-language models.

- **SoundMind RL Algorithm:** Taking into account the unique challenges of audio reasoning generation, we design a strict format-based reward mechanism that maintains reasoning coherence across modalities. By combining the improved REINFORCE++ algorithm with the reward design principles of Deepseek-R1, we fine-tune Qwen2.5-Omni-7B using reinforcement learning, achieving state-of-the-art performance across three types of modality input-output combinations on the ALR test set.

2 Related Work

Open-Source Datasets for Audio Reasoning. Currently, there are very few publicly available datasets for audio reasoning. The audio datasets released in recent years typically use audio as a descriptive input, combined with text, and are fed jointly into ALMs. However, the final annotations in these datasets are purely textual (Suzgun et al., 2023; Kong et al., 2024; Ghosh et al., 2025). Recently, some researchers have begun to recognize the critical role of CoT prompting in audio reasoning datasets (Xie et al., 2025b). Nevertheless, our study is the first open-source dataset in which audio serves as the annotated modality for reasoning.

Chain-of-Thought Reasoning. LLMs enhance their reasoning ability through in-context learning (ICL), which processes prompts within the surrounding context. CoT techniques further reinforce this capability. Prominent CoT approaches include Tree-of-Thought (ToT) (Yao et al., 2023), manually crafted few-shot CoT (Wei et al., 2022), and various automatic generation strategies (Jin et al., 2024). Recent works have also examined the necessity, theoretical underpinnings, and task-specific effectiveness of CoT reasoning (Sprague et al., 2025). OpenAI's ol model (Jaech et al., 2024) has reignited interest in CoT prompting and is often paired with reinforcement learning-based training

Datasets	**Type**	**Transcripts**	**CoT annotations**	**Samples**	**Hours**
CoTA	Sound, Speech, Music	✘	✔	1.2M	6K
AudioSkills	Speech	✘	✘	4.2M	9.3K
LongAudio	Audio	✘	✘	263K	8.5K
Big Bench Audio	Speech	✘	✘	1000	2
ALR (Our)	Speech	✔	✔	6446	1K

Table 1: Comparison of reasoning audio datasets. ALR is the only dataset providing both transcripts and chain-of-thought annotations in addition to audio, supporting step-by-step reasoning and multimodal evaluation.

approaches (Hu, 2025; Xie et al., 2025a).

Multimodal Chain-of-Thought Reasoning. CoT prompting has seen notable advancements in the multimodal domain. For instance, Visual-CoT (Shao et al., 2024) integrates object detection to assist reasoning, while LLaVA-CoT (Xu et al., 2024) and MAmmoTH-VL (Guo et al., 2024) improve performance via dataset augmentation. However, CoT applications in the audio domain are still in their infancy. Audio-CoT (Ma et al., 2025) demonstrates that zero-shot CoT prompting yields improvements on simple audio tasks, but remains inadequate for complex reasoning.

Although current ALMs have made progress in comprehension and real-time response, their capability in CoT-style reasoning remains underexplored. Our study addresses this research gap by systematically investigating the application of CoT techniques within ALMs.

3 Audio Logic Reasoning Dataset

3.1 Main Features

We introduce the **Audio Logic Reasoning (ALR)** dataset, a novel benchmark specifically designed to advance audio-based reasoning research. Unlike most existing datasets that rely solely on textual annotations, ALR provides both audio-level and CoT annotations, enabling models to learn and reason directly from auditory inputs. The dataset focuses on speech-based scenarios, aligning with real-world applications such as spoken dialogue understanding and auditory commonsense reasoning.

As shown in Table 1, existing audio datasets such as CoTA (Xie et al., 2025b), AudioSkills (Ghosh et al., 2025), LongAudio (Ghosh et al., 2025), and Big Bench Audio (Suzgun et al., 2023) either lack audio-level annotations or do not provide CoT reasoning traces. For instance, while CoTA includes CoT-style annotations, it does not offer audio annotations/transcripts, limiting its ability to support models that operate directly on audio. Similarly, AudioSkills and LongAudio provide a large volume of audio samples but lack both reasoning annotations and step-by-step rationales.

In contrast, ALR is the only dataset in this comparison that includes both audio and CoT annotations. Although it contains fewer samples (6,446) than large-scale datasets such as AudioSkills, it is more focused on logic-oriented reasoning tasks and offers 1,074 hours of carefully annotated speech audio. This dual-modality annotation makes ALR a uniquely valuable resource for training and evaluating Logic-Aware ALMs in tasks requiring structured reasoning grounded in audio perception.

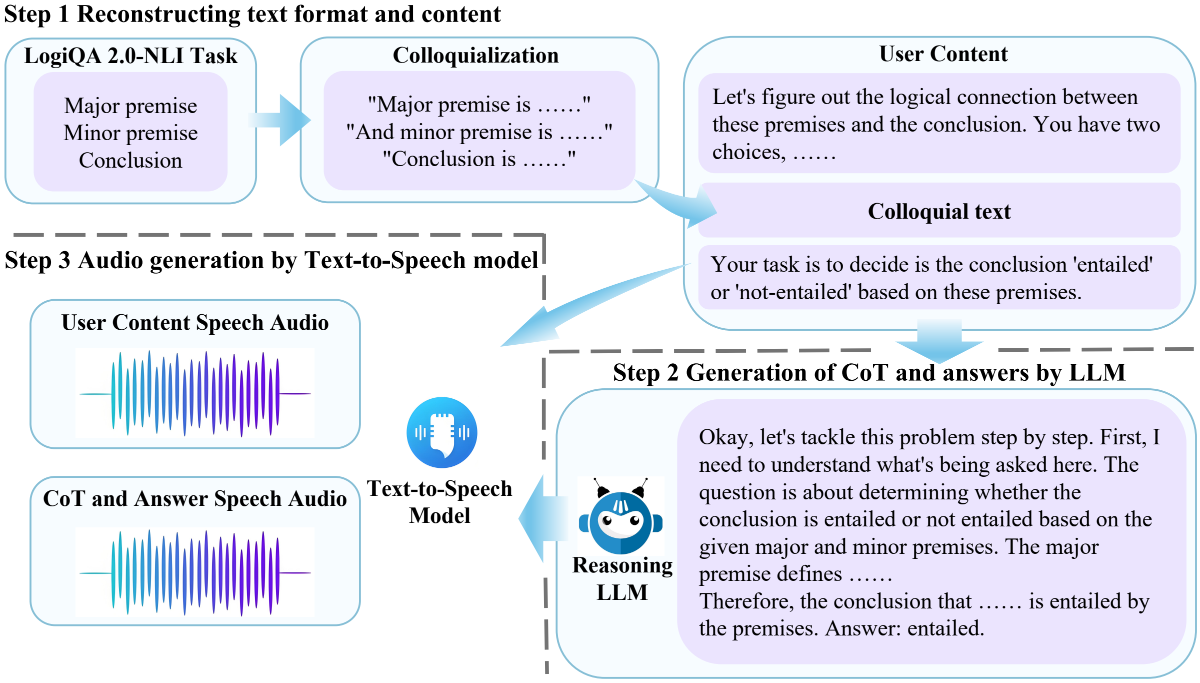
3.2 Data Generation Pipeline

The Audio Logic Reasoning (ALR) dataset is constructed via a structured pipeline that converts textual logical reasoning tasks into natural spoken audio interactions, enabling audio-based reasoning model development. The full process is in Figure 2.

We begin with logic problems from the LogiQA 2.0 NLI (Liu et al., 2023), which provides a structured triplet: a major premise, a minor premise, and a conclusion. These inputs are first processed through a **colloquialization** module to convert formal logic into natural, conversational language (e.g., *Major premise is . . .* , *Let's figure out the logical connection. . .*). This step enhances the naturalness of the audio and simulates real user prompts.

The colloquialized content and instructions are combined into **user content**, which is then synthesized into **speech audio** using the high-quality text-to-speech model MegaTTS 3 (Jiang et al., 2025). This forms the input side of the dataset. The prompts we used can be found in Table 2.

To generate the output, we use the large language model DeepSeek-R1 to produce CoT reasoning and final answers. These reasoning traces provide step-by-step explanations and improve supervision quality. The CoT outputs are also synthesized into


Step 1 Reconstructing text format and content
LogiQA 2.0-NLI Task
Major premise
Minor premise
Conclusion
Colloquialization
User Content
Colloquial text
Step 3 Audio generation by Text-to-Speech model
User Content Speech Audio
CoT and Answer Speech Audio
Text-to-Speech Model
Step 2 Generation of CoT and answers by LLM
Reasoning LLM

"Major premise is ……"
"And minor premise is ……"
"Conclusion is ……"

Let's figure out the logical connection between these premises and the conclusion. You have two choices, ……

Your task is to decide is the conclusion 'entailed' or 'not-entailed' based on these premises.

Okay, let's tackle this problem step by step. First, I need to understand what's being asked here. The question is about determining whether the conclusion is entailed or not entailed based on the given major and minor premises. The major premise defines ……
Therefore, the conclusion that …… is entailed by the premises. Answer: entailed.

Figure 2: Three-stage pipeline for constructing ALR samples. Step 1: Structured logical inputs are converted into natural prompts through colloquialization. Step 2: A large language model generates chain-of-thought reasoning and final answers. Step 3: Both input and output texts are synthesized into speech using a text-to-speech model, producing paired audio data aligned with logical reasoning.

audio using MegaTTS 3, producing the corresponding spoken response.

As a result, each ALR sample consists of a pair of spoken audio segments: one representing the user query and the other containing the reasoning and answer, as shown in Figure 3.

4 SoundMind Algorithm

We improve our system through iterative optimization of rule-based rewards. All λ in the following equations are hyperparameters.

Answer Format Correctness Evaluation. To ensure the correctness of answer formatting, we first require that the token “Answer:” must appear within the last five characters of the model response. Considering the dual-modality annotation characteristics of the ALR dataset, we design two specific format scoring methods (denoted as S_{format}), calculated as follows:

$$S_{\text{format}}^{(1)} = \begin{cases} \lambda_1, & \text{if correct for text tokens} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

$$S_{\text{format}}^{(2)} = \begin{cases} \lambda_2, & \text{if correct for audio tokens} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Answer Correctness Evaluation. Once the format compliance is verified, this module evaluates the factual accuracy of the model response. Specifically, the answer score (S_{answer}) is computed based on the consistency between the model's predicted answer and the ground truth answer, using the following formulation:

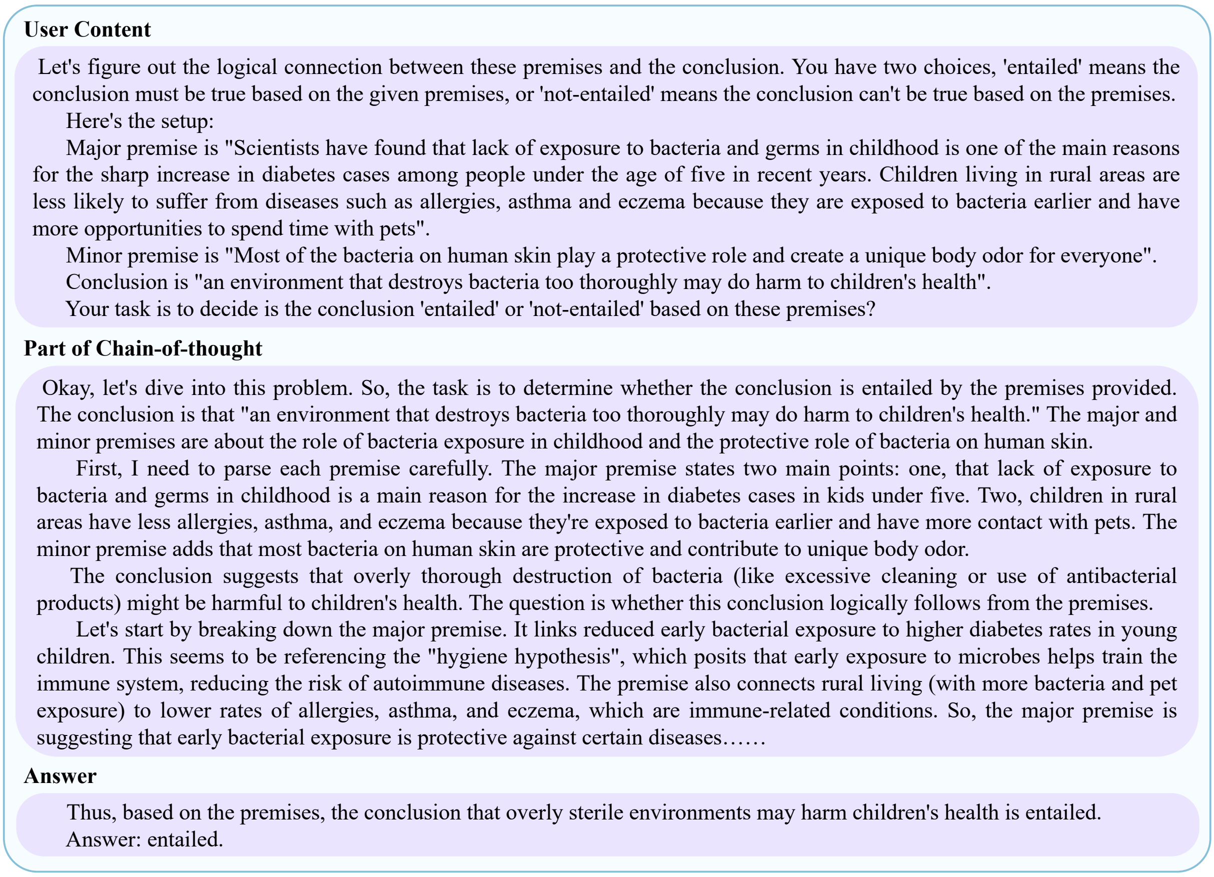
$$S_{\text{answer}} = \begin{cases} \lambda_3, & \text{if answer = ground truth} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Reasoning Length Evaluation. We additionally introduce a reward evaluation based on the reasoning length of the model response. This score (S_{len}) is computed by comparing the ratio of the model output length to the reference reasoning length. In light of the ALR dataset's dual-modality annotations, we design two distinct length evaluation methods, defined as follows:

$$S_{\text{len}}^{(1)} = \lambda_4 \times \min\left(1, \frac{L_{\text{model}}}{L_{\text{annotation}}}\right), \tag{4}$$

$$S_{\text{len}}^{(2)} = \lambda_5 \times \min\left(1, \frac{T_{\text{model}}}{T_{\text{annotation}}}\right). \tag{5}$$

L denotes the length of text tokens, and T denotes the length of audio tokens.



User Content

Let's figure out the logical connection between these premises and the conclusion. You have two choices, 'entailed' means the conclusion must be true based on the given premises, or 'not-entailed' means the conclusion can't be true based on the premises.

Here's the setup:

Major premise is "Scientists have found that lack of exposure to bacteria and germs in childhood is one of the main reasons for the sharp increase in diabetes cases among people under the age of five in recent years. Children living in rural areas are less likely to suffer from diseases such as allergies, asthma and eczema because they are exposed to bacteria earlier and have more opportunities to spend time with pets".

Minor premise is "Most of the bacteria on human skin play a protective role and create a unique body odor for everyone".

Conclusion is "an environment that destroys bacteria too thoroughly may do harm to children's health".

Your task is to decide is the conclusion 'entailed' or 'not-entailed' based on these premises?

Part of Chain-of-thought

Okay, let's dive into this problem. So, the task is to determine whether the conclusion is entailed by the premises provided. The conclusion is that "an environment that destroys bacteria too thoroughly may do harm to children's health." The major and minor premises are about the role of bacteria exposure in childhood and the protective role of bacteria on human skin.

First, I need to parse each premise carefully. The major premise states two main points: one, that lack of exposure to bacteria and germs in childhood is a main reason for the increase in diabetes cases in kids under five. Two, children in rural areas have less allergies, asthma, and eczema because they're exposed to bacteria earlier and have more contact with pets. The minor premise adds that most bacteria on human skin are protective and contribute to unique body odor.

The conclusion suggests that overly thorough destruction of bacteria (like excessive cleaning or use of antibacterial products) might be harmful to children's health. The question is whether this conclusion logically follows from the premises.

Let's start by breaking down the major premise. It links reduced early bacterial exposure to higher diabetes rates in young children. This seems to be referencing the "hygiene hypothesis", which posits that early exposure to microbes helps train the immune system, reducing the risk of autoimmune diseases. The premise also connects rural living (with more bacteria and pet exposure) to lower rates of allergies, asthma, and eczema, which are immune-related conditions. So, the major premise is suggesting that early bacterial exposure is protective against certain diseases……

Answer

Thus, based on the premises, the conclusion that overly sterile environments may harm children's health is entailed.

Answer: entailed.

Figure 3: An example in text from the ALR dataset. The sample consists of three components: (1) **User Content**, which includes a natural-language prompt and a structured logical triplet (major premise, minor premise, and conclusion); (2) **Chain-of-Thought** response generated by a large language model, which explains the logical reasoning behind the entailment decision; and (3) the final **Answer**. All components are available in both text and synthesized audio formats, supporting multimodal training and evaluation.

REINFORCE++ Policy Optimization. In our setting, the policy π_θ corresponds to a large-scale ALM, which receives an audio question as input and produces a reasoning response in either text or audio form. To optimize the model with the above composite reward, we adopt the REINFORCE++ (Hu, 2025)—a clipped policy-gradient method that eliminates the need for a value (critic) network while leveraging PPO-style stability and sample efficiency. Specifically, REINFORCE++ updates the policy by maximizing the following objective:

$$\mathcal{J}_{\text{REINFORCE++}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D},\, o_{\le t}\sim\pi_{\theta_{\text{old}}}} \left[\min\left(r_t(\theta)\, \hat{A}_t,\ \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\ \hat{A}_t \right) \right] \tag{6}$$

where x is the audio input, y is the generated response, and $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the importance sampling ratio. $\hat{A}_t$ is the normalized advantage, given by: $A_t = R(x,y) - \beta \sum_{i=t}^{T} \text{KL}(i)$, $\hat{A}_t = \frac{A_t - \mu_A}{\sigma_A}$. Here, $R(x,y)$ is the cumulative reward, composed of the weighted sum of the above reward terms (S_{format}, S_{answer}, S_{len}), and $\text{KL}(i)$ denotes the token-level KL divergence between the current policy and the reference SFT model. The KL term is defined as:

$$\text{KL}(i) = \log \frac{\pi_\theta(a_i|s_i)}{\pi_{\text{SFT}}(a_i|s_i)}, \tag{7}$$

where π_θ is the current policy being optimized, π_{SFT} is the fixed reference (supervised fine-tuned) model, β is the KL penalty coefficient, and μ_A, σ_A are the batch mean and standard deviation for normalization. ϵ is the PPO clip parameter. REINFORCE++ combines the stability of PPO's clipped surrogate loss with the efficiency of critic-free Monte Carlo policy gradient updates, using a token-level KL penalty and batch-normalized advantages. It trains audio-language models for end-to-end reasoning over audio questions.

Position	Prompt
System	Your task is to decide if the conclusion is "entailed" or "not-entailed" based on these premises. You are a wise person who answers two-choice questions, "entailed" or "not entailed". Use plain text for thought processes and answers, not markdown or LaTeX. The thought process and response style should be colloquial, which I can then translate directly into audio using the TTS model. The final output is the Answer, nothing else, and the format is Answer: YOUR ANSWER. For example: "Answer: entailed." or "Answer: not entailed." The final answer must contain nothing else! The thought process should be very complete, careful, and cautious. When you think and generate a chain of thought, you need to test your answer from various angles.
Before the major premise	Let's figure out the logical connection between these premises and the conclusion. You have two choices: "entailed" means the conclusion must be true based on the given premises, or "not-entailed" means the conclusion can't be true based on the premises. Here's the setup:
Behind the conclusion	Your task is to decide is the conclusion is "entailed" or "not-entailed" based on these premises.

Table 2: Instructional prompts at different positions during chain-of-thought generation. The *system* prompt defines the overall task, output format, and desired reasoning style for speech synthesis. Contextual prompts inserted before the major premise and after the conclusion further guide the model in interpreting logical relationships.

Split	# Entailed	# Not-ent.	Avg. Inp. Tok.	Avg. Out. Tok.	Avg. Inp. Dur. (s)	Avg. Out. Dur. (s)
Training	2326	2858	182	1683	62.33	608.42
Test	296	360	158	1424	57.90	586.51
Validation	264	342	155	1426	57.85	586.39

Table 3: Detailed statistics of the Audio Logic Reasoning (ALR) dataset. "# Entailed" and "# Not-ent." indicate the number of entailed and not-entailed samples, respectively. "Avg. Inp./Out. Tok." denotes the average input/output tokens. "Avg. Inp./Out. Dur. (s)" indicates the average duration (in seconds) of input/output sequences.

5 Experiments

5.1 Setup

We fine-tune the Qwen2.5-Omni-7B model on the ALR dataset using the proposed SoundMind algorithm. All experiments were conducted under a consistent hardware environment, consisting of an Intel(R) Xeon(R) Platinum 8468 CPU and 8 NVIDIA H800 GPUs, each equipped with 80 GB of memory. The weighting coefficients for the reward components were set as follows: $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 2.0$, $\lambda_4 = 1.0$, and $\lambda_5 = 0.75$. The training procedure was executed for 50,000 steps. All other hyperparameter settings followed those used in Logic-RL (Xie et al., 2025a).

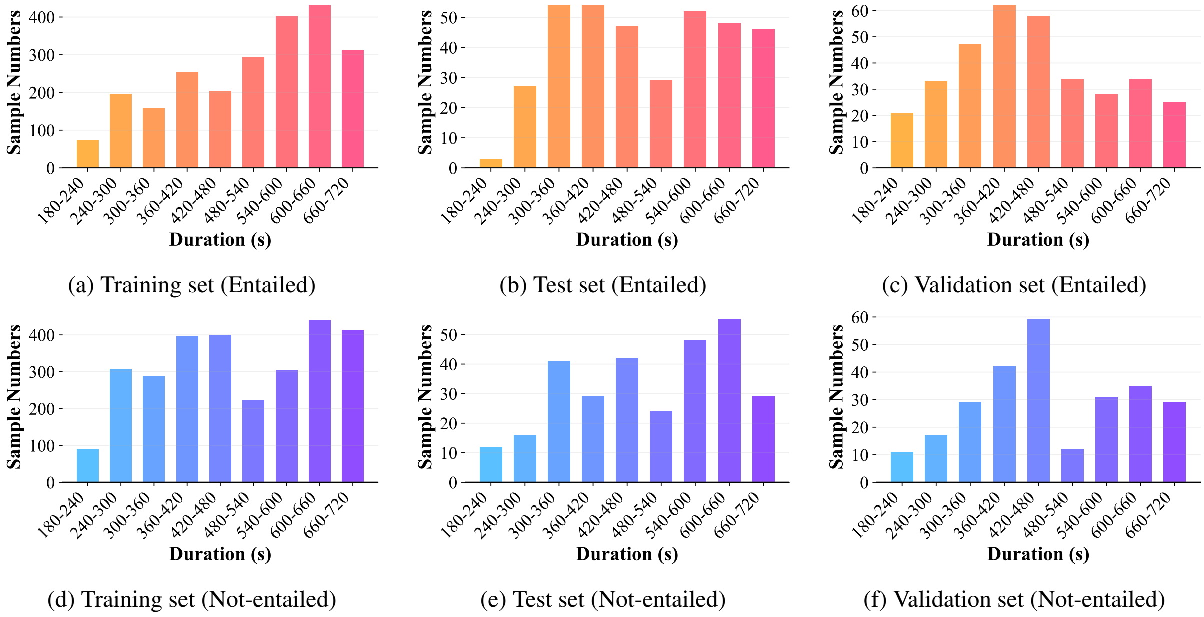
To consider the dual-modality nature of both input and output in the ALR dataset, we conduct experiments across three settings: ① Table 4: audio-only input with text-only reasoning output. We select five multimodal LLMs as baselines: MiniCPM-o (Yao et al., 2024), Gemini-Pro-V1.5 (Team et al., 2024), Baichuan-Omni-1.5 (Li et al., 2025), Qwen2-Audio (Chu et al., 2024), and Qwen2.5-Omni-7B (Xu et al., 2025). ② Table 5: text-only input with audio-only reasoning output. Since there are few bimodal LLMs with a similar number of parameters to Qwen2.5-Omni-7B, we use Qwen2.5-Omni-7B as the baseline. ③ Table 6: audio-only input with audio-only reasoning output.

5.2 Dataset Analysis

Table 3 presents statistics of the ALR dataset, highlighting its scale, balance, and linguistic richness across training, validation, and test splits. The dataset comprises a total of 6,446 samples, with a relatively balanced distribution between entailed and not-entailed instances, ensuring fair coverage of both classes for entailment prediction.

The average length of user content is approximately 160–180 tokens across all splits, while the generated CoT responses are significantly longer, averaging over 1,400 tokens. This discrepancy underscores the dataset's emphasis on step-by-step, explainable reasoning, which is crucial for supporting interpretability in audio-based reasoning tasks.

In terms of audio representation, the user content audio averages around one minute in duration, whereas the CoT reasoning speech spans nearly 10 minutes per sample. This extended audio format provides an excellent benchmark for evaluating models' long-form audio understanding and reasoning capabilities.



(a) Training set (Entailed)

(b) Test set (Entailed)

(c) Validation set (Entailed)

(d) Training set (Not-entailed)

(e) Test set (Not-entailed)

(f) Validation set (Not-entailed)

Figure 4: Histograms of audio durations across dataset splits and entailment labels. Subfigures (a–c) correspond to the training, test, and validation sets for "Entailed" samples; (d–f) show the same splits for "Not-entailed" samples.

As shown in Figure 4, the ALR dataset exhibits a wide range of duration distributions that reflect critical considerations for audio-logical reasoning research. Audio segments predominantly span 3–12 minutes (with 28.6% of training samples in the 540–720s range), balancing sufficient contextual information with manageable computational complexity. The design deliberately avoids shorter clips lacking substantive reasoning content.

The dataset demonstrates a modest class imbalance (44.9% entailed vs. 55.1% not-entailed overall), mirroring real-world reasoning scenarios. This imbalance is particularly pronounced in mid-range durations (240–420s), where audio complexity appears to challenge entailment determination. Notably, the 480–540s bin in the training set shows an inverse distribution (56.9% entailed), possibly reflecting unique acoustic features or annotation artifacts.

Strategic split allocation maintains consistent entailment ratios across training (42.8%), test (54.9%), and validation (56.3%) sets, ensuring reliable evaluation while preserving inherent complexity. Proportional representation across all duration bins in each split helps prevent duration-based bias. The near-balanced nature of the test set (360 entailed vs. 296 not-entailed) supports rigorous and fair model assessment.

Overall, this distribution profile directly supports the dataset's objective of advancing bimodal reasoning by providing diverse yet controlled audio-text interaction scenarios. These settings challenge models to develop genuine cross-modal understanding, rather than relying on duration-specific pattern recognition.

5.3 Results and Analysis

Model	Accuracy (%)↑
MiniCPM-o	73.17
Gemini-Pro-V1.5	74.54
Baichuan-Omni-1.5	70.58
Qwen2-Audio	58.23
Qwen2.5-Omni-7B	77.59
Qwen2.5-Omni-7B-RL (Our)	**81.40**

Table 4: Test accuracy (%) of different models on the ALR benchmark for audio-to-text logical reasoning (audio input, text output).

SoundMind achieves state-of-the-art performance on the audio-to-text reasoning task. Table 4 presents results for generating textual reasoning outputs from audio-only inputs. Our Qwen2.5-Omni-7B-RL model achieves an accuracy of 81.40%, outperforming all evaluated baselines. Compared to its supervised counterpart Qwen2.5-Omni-7B (77.59%), the reinforcement-tuned variant achieves a notable absolute gain of 3.81%. This indicates that the SoundMind reward

Model	WER (%)↓	Acc. (%)↑
Qwen2.5-Omni-7B	**2.18**	80.79
Qwen2.5-Omni-7B-RL (Our)	6.99	**83.84**

Table 5: Test set performance (% WER↓, accuracy↑) of models on the ALR benchmark for text-to-audio reasoning (text input, speech output).

Model	WER (%)↓	Acc. (%)↑
Qwen2.5-Omni-7B	**2.23**	77.59
Qwen2.5-Omni-7B-RL (Our)	8.95	**81.40**

Table 6: Test set performance (% WER↓, accuracy↑) of models on the ALR benchmark for audio-to-audio reasoning (audio input, speech output).

framework significantly improves the model's ability to generate accurate, text-based logical conclusions based on audio inputs. Among non-RL baselines, Gemini-Pro-V1.5 achieves the second-best performance with 74.54%, followed by MiniCPM-o (73.17%), and Baichuan-Omni-1.5 (70.58%). Qwen2-Audio lags behind with 58.23%, highlighting that generic multimodal alignment is insufficient for reasoning-heavy tasks without dedicated optimization. Qwen2.5-Omni-7B-RL improves upon Gemini-Pro by a margin of 6.86%, and surpasses MiniCPM-o by 8.23%, establishing a clear lead across models of comparable or larger capacity. These results validate that our reinforcement learning setup improves reasoning accuracy and generalizes effectively across audio entailment tasks, driven by structured reward signals for answer format, answer correctness, and reasoning length.

SoundMind enhances logical structuring in text-to-audio reasoning despite modality shift. In Table 5, we evaluate the model's ability to generate speech-based reasoning from text-only prompts. The RL-enhanced model achieves 83.84% accuracy, representing a 3.05% improvement over the supervised baseline. This result highlights that even when the input lacks audio grounding, the reward structure guides the model toward more structured and consistent reasoning in speech. The increased reasoning depth, induced by chain-of-thought reinforcement signals, indicates that SoundMind is capable of refining response quality across modalities—not only in interpreting audio, but also in generating semantically coherent outputs in speech form.

In fully speech-based settings, SoundMind generalizes reasoning without textual anchors. Table 6 investigates the most demanding configuration: generating spoken reasoning from audio-only inputs, with no access to textual content during inference. In this dual-audio scenario, our model maintains a strong performance gain, raising accuracy from 77.59% to 81.40%. This demonstrates that SoundMind can internalize abstract logical dependencies in audio alone—without relying on textual shortcuts. Such robustness is especially valuable for real-world applications involving spoken dialogue systems, voice-based tutoring agents, or accessibility tools that require full audio-audio reasoning capabilities.

Across all tasks, SoundMind consistently improves logical accuracy while maintaining acceptable speech quality. Our Qwen2.5-Omni-7B-RL model outperforms all baselines in reasoning accuracy across the three evaluated input-output modality pairs. These results confirm that reward-guided reinforcement learning can strengthen logical inference capabilities in audio-language models. While we observe a moderate rise in Word Error Rate (WER) in speech-based outputs, this trade-off is expected given the model's increased focus on semantic correctness and structured explanation. The observed WER remains within acceptable bounds and does not hinder interpretability. Future work can explore fluency-aware RL objectives or dual-path reward balancing to further harmonize reasoning fidelity and audio naturalness.

6 Conclusion

We introduce **SoundMind**, a novel rule-based reinforcement learning framework that empowers large-scale audio-language models with advanced logical reasoning capabilities across both audio and textual modalities. To enable such training, we build the **Audio Logical Reasoning** (ALR) dataset, a dual-modality benchmark comprising 6,446 high-quality samples annotated with chain-of-thought reasoning in both audio and text forms. Experimental results demonstrate that our method significantly improves performance and establishes new state-of-the-art results on the ALR benchmark, consistently outperforming strong baselines across three reasoning settings: text-to-audio, audio-to-text, and audio-to-audio. We hope this work can serve as a foundation for future research in logic-oriented large-scale audio-language modeling and in reinforcement-driven multimodal reasoning.

Limitations

While our approach demonstrates strong performance, certain aspects warrant further exploration. In particular, the Word Error Rate (WER) in audio generation remains competitive, but still leaves room for improving reasoning fidelity during speech synthesis. Addressing this limitation could lead to more robust, adaptive, and human-aligned audio reasoning systems in future work.

Ethical Considerations

We have not identified any ethical concerns directly related to this study.

Acknowledgment

This study is supported by the Department of Defense grant HT9425-23-1-0267.

References

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*.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*.

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*.

Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*.

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*.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, et al. 2025. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*.

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*.

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*.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. 2025. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *International Conference on Learning Representations*.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *Transactions on Audio, Speech, and Language Processing*.

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

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Workshop on Reasoning and Planning for Large Language Models at ICLR2025*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL*.

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*.

Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2024. Alphazero-like tree-search can guide large language model decoding and training. In *International Conference on Machine Learning*.

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*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025a. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025b. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.

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*.

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

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*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

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