Step-Audio-R1 Technical Report

StepFun-Audio Team

StepAudio R1 Official Github Page

StepAudio R1 Official Demo Page

Abstract

Recent advances in reasoning models have demonstrated remarkable success in text and vision domains through extended chain-of-thought deliberation. However, a perplexing phenomenon persists in audio language models: they consistently perform better with minimal or no reasoning, raising a fundamental question—**can audio intelligence truly benefit from deliberate thinking?** We introduce Step-Audio-R1, the first audio reasoning model that successfully unlocks reasoning capabilities in the audio domain. Through our proposed Modality-Grounded Reasoning Distillation (MGRD) framework, Step-Audio-R1 learns to generate audio-relevant reasoning chains that genuinely ground themselves in acoustic features rather than hallucinating disconnected deliberations. Our model exhibits strong audio reasoning capabilities, surpassing Gemini 2.5 Pro and achieving performance comparable to the state-of-the-art Gemini 3 Pro across comprehensive audio understanding and reasoning benchmarks spanning speech, environmental sounds, and music. These results demonstrate that reasoning is a transferable capability across modalities when appropriately anchored, transforming extended deliberation from a liability into a powerful asset for audio intelligence. By establishing the first successful audio reasoning model, Step-Audio-R1 opens new pathways toward building truly multimodal reasoning systems that think deeply across all sensory modalities.

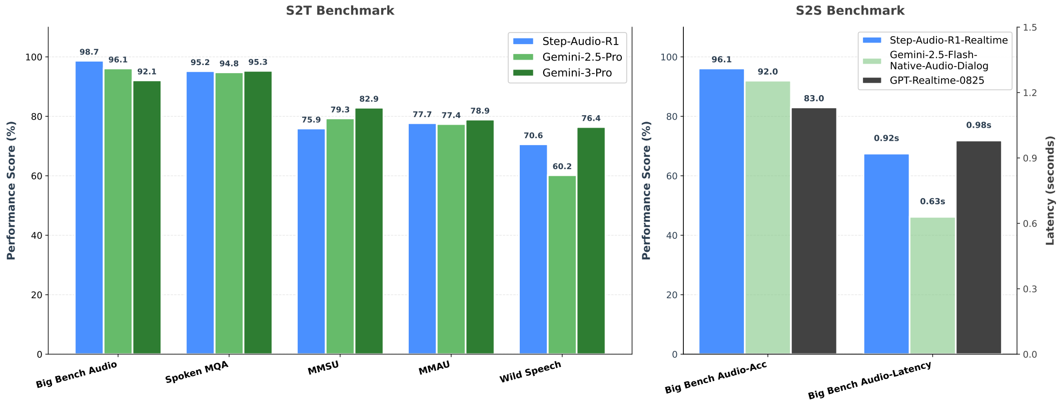


Figure 1: Benchmark performance of Step-Audio-R1

1 Introduction

Chain-of-thought reasoning has transformed modern artificial intelligence, enabling language models to solve complex mathematical problems [1–3], generate executable code [4], and engage in sophisticated logical deduction through extended deliberation [5]. Vision-language models have similarly adopted this paradigm, leveraging deliberate reasoning to interpret spatial relationships [2], analyze visual scenes, and answer intricate questions about images [6, 7]. Underlying these successes is a fundamental principle known as test-time compute scaling: allocating more computational resources during inference—through longer chains of thought, iterative refinement, or search—predictably improves model performance [8–10]. This scaling law has become so robust that it now guides the design and deployment of AI systems across modalities.

The audio domain, however, presents a stark exception to this principle. Existing audio language models consistently demonstrate superior performance with minimal or no reasoning [11, 12]. Empirical observations across benchmarks reveal that direct responses outperform elaborate chain-of-thought explanations, with performance systematically degrading as reasoning length increases [11, 13, 14]. This inverted scaling behavior persists across architectures, training methodologies, and model scales [12, 13], suggesting a fundamental incompatibility between test-time compute scaling and auditory intelligence. This raises a critical question:

Is audio inherently resistant to deliberate reasoning?

Recent efforts have attempted to address this anomaly through reinforcement learning approaches that employ language model judges to verify consistency between reasoning chains and final answers [15, 16]. While these methods improve alignment, they treat the symptom rather than the root cause—enforcing consistency without understanding *why* reasoning fails in audio. Through systematic case studies, we uncover a striking pattern: existing audio language models engage in *textual surrogate reasoning* rather than acoustic reasoning. When prompted to deliberate, models systematically reason from the perspective of transcripts or textual captions instead of acoustic properties—for instance, attributing musical melancholy to "lyrics mentioning sadness" rather than "minor key progressions and descending melodic contours". This leads to a critical hypothesis: **the performance degradation stems not from reasoning itself, but from reasoning about the wrong modality**. We trace this to a fundamental design choice: most audio language models initialize their reasoning capabilities through supervised fine-tuning on COT [17] data derived from text-based models [12, 13]. Consequently, these models inherit linguistic grounding mechanisms, creating a modality mismatch that undermines performance as reasoning chains lengthen.

To validate this hypothesis and unlock reasoning capabilities in audio, we propose *Modality-Grounded Reasoning Distillation* (MGRD), an iterative training framework that progressively shifts reasoning from textual abstractions to acoustic properties. Starting from text-based reasoning initialization, MGRD employs iterative cycles of self-distillation and refinement on audio tasks, systematically curating reasoning chains that genuinely ground in acoustic analysis. Through these iterations, we obtain Step-Audio-R1, the first audio reasoning model that successfully benefits from test-time compute scaling, outperforming Gemini 2.5 Pro [14] and demonstrating capabilities competitive with the latest Gemini 3 Pro [18] across comprehensive audio benchmarks. These results confirm that reasoning is a transferable capability across modalities when appropriately anchored, transforming extended deliberation from a liability into a powerful asset for audio intelligence.

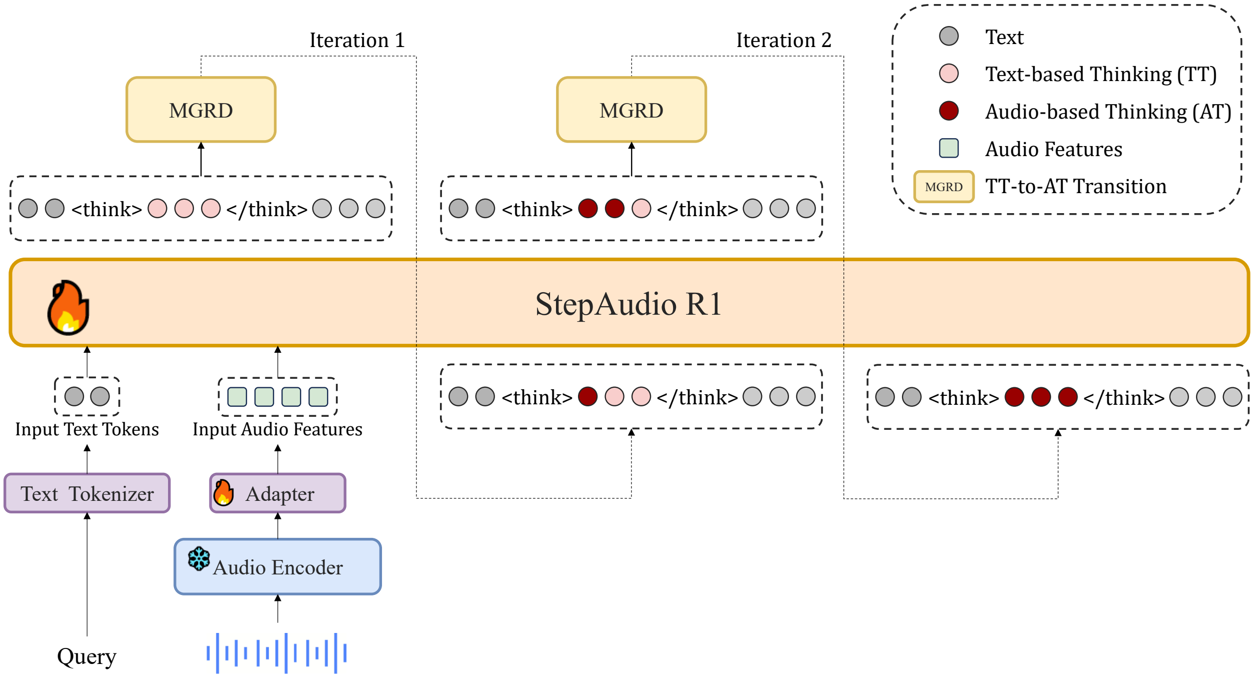


Figure 2: The overview of Step-Audio-R1

2 Model Overview

Drawing from the architecture of our previous Step-Audio 2 [13], Step-Audio-R1 is designed for audio-based reasoning tasks. As shown in Figure 2, the model consists of an audio encoder, an audio adaptor, and an LLM decoder.

For the audio encoder, we utilize the Qwen2 audio encoder [19], which is pretrained on various speech and audio understanding tasks. The audio encoder has an output frame rate of 25 Hz and is frozen during the entire training process. An audio adaptor with a downsampling rate of 2, identical to the one in Step-Audio 2, is employed to connect the audio encoder to the LLM, thereby reducing the output frame rate to 12.5 Hz.

The LLM decoder, based on Qwen2.5 32B [20], directly takes the latent audio features from the audio adaptor as input to generate a purely textual output. The model is structured to first generate the reasoning content, followed by the final reply.

A key innovation in this process is the Modality-Grounded Reasoning Distillation (MGRD) method. Initially, the model's reasoning process may operate on a purely semantic level. MGRD iteratively refines these thoughts, progressively strengthening their connection to the underlying audio features until they evolve into "native audio think." This distillation process ensures that the model's reasoning is not merely about the transcribed text, but is deeply grounded in the acoustic nuances of the audio itself, leading to a more holistic and accurate final response.

Step-Audio-R1 is pretrained using the same data and methodology as Step-Audio 2. Following this, the model undergoes a Post-Training phase, with specific details provided in Section 4.

3 Data Preparation

3.1 Data for Cold-Start

Our cold-start phase is designed to jointly elicit audio reasoning capabilities through a combination of Supervised Fine-Tuning (SFT) and Reinforcement Learning with Verified Reward (RLVR). This phase utilizes a total dataset of 5 million samples. This token budget is comprised of 1B tokens from text-only data, with the remaining 4B tokens derived from our audio-side data.

The data types are as follows:

- **Audio Data:** This includes Automatic Speech Recognition (ASR), Paralinguistic Understanding, and standard Audio Question Text Answer (AQTA) dialogues.
- **Audio CoT Data:** We incorporate AQTA Chain-of-Thought (CoT) data, which is generated via self-distillation from our own model after its audio reasoning capabilities were elicited. This CoT data constitutes 10% of our total audio dataset.
- **Text Data:** This includes text-only dialogues (in both single-turn and multi-turn formats) covering topics such as knowledge-based QA, novel continuation, role-playing, general chat, and emotional conversations. It also incorporates the text CoT data, which focuses on math and code.

A critical aspect of our data strategy is the standardized reasoning format. To train the model to recognize the reasoning structure, we prepend all samples lacking native CoT with an empty `<think>` tag. The format is standardized as: `<think>\n\n</think>\n{response}`

3.2 Data for RL

For the subsequent Reinforcement Learning (RL) phase, we curated a smaller, high-quality dataset of 5,000 samples. This dataset is composed of 2,000 high-quality text-only samples (focusing on math and code) and 3,000 augmented speech-based QA samples. The augmentation methods used to process this data are described in detail in Section 4.2.

4 Post-Training Recipes

4.1 Foundation Training: Reasoning Initialization and Format Alignment

We establish fundamental reasoning capabilities through a two-stage training process that builds robust reasoning primitives while maintaining basic audio understanding.

Supervised Chain-of-Thought Initialization. Given a base audio-language model π_{θ_0}, we perform supervised fine-tuning on chain-of-thought demonstrations from both task-oriented and conversational domains, along with audio data to preserve multimodal capabilities. The training objective unifies three data sources:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{(q,r,a)\sim\mathcal{D}_{\text{task}}}\left[\log \pi_\theta(r, a \mid q)\right] + \mathbb{E}_{(c,r,s)\sim\mathcal{D}_{\text{conv}}}\left[\log \pi_\theta(r, s \mid c)\right] + \mathbb{E}_{(x_{\text{audio}},q,a)\sim\mathcal{D}_{\text{audio}}}\left[\log \pi_\theta(a \mid x_{\text{audio}}, q)\right] \tag{1}$$

where (q, r, a) denotes task questions with reasoning chains and answers, (c, r, s) represents conversational contexts with deliberation and responses, and (x_{audio}, q, a) indicates audio questions with

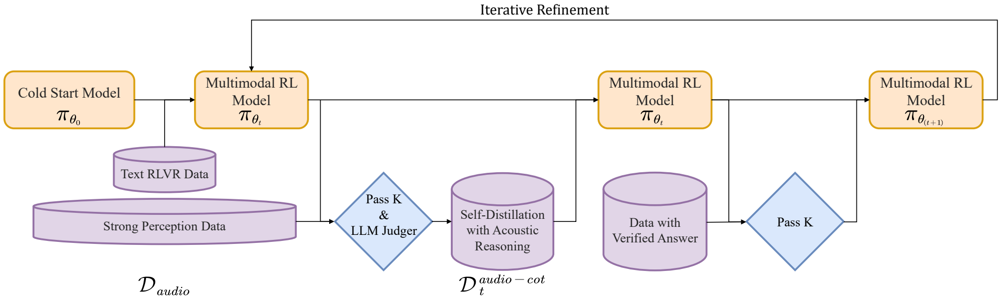

Iterative Refinement
Cold Start Model π_{θ_0}
Multimodal RL Model π_{θ_t}
Text RLVR Data
Strong Perception Data
$\mathcal{D}_{audio}$
Pass K & LLM Judger
Self-Distillation with Acoustic Reasoning
$\mathcal{D}_t^{audio-cot}$
Multimodal RL Model π_{θ_t}
Data with Verified Answer
Pass K
Multimodal RL Model $\pi_{\theta_{(t+1)}}$

Figure 3: Modality-Grounded Reasoning Distillation

direct answers. For audio data, we use empty reasoning markers (i.e., <think>\n\n</think>\n) to maintain the structural format without actual deliberation content. This tri-modal training instills diverse reasoning patterns in text domains—spanning analytical problem-solving, code generation, logical inference, and contextual dialogue—while preserving the model's audio understanding capabilities for subsequent acoustic reasoning distillation.

Reinforcement Learning with Verified Rewards. Building upon the supervised foundation, we refine reasoning quality on task-oriented data through Reinforcement Learning with Verified Rewards (RLVR) [3, 21]. For mathematical problems, coding challenges, and logical puzzles, the model samples reasoning trajectories and receives binary verification rewards:

$$R(r, a) = \begin{cases} 1, & \text{if } a = a^* \\ 0, & \text{else} \end{cases} \tag{2}$$

We optimize using Proximal Policy Optimization [22] without KL penalty constraints, maximizing expected reward:

$$\mathcal{L}_{\text{RLVR}} = \mathbb{E}_{\mathcal{D}_{\text{task}}} \left[R(r, a) \right] \tag{3}$$

This allows free exploration of reasoning strategies while maintaining answer accuracy through outcome-based verification.

4.2 Modality-Grounded Reasoning Distillation

With textual reasoning foundation established, we now address the core challenge: transforming reasoning capabilities from textual abstractions to acoustic grounding. We propose an iterative self-distillation framework that progressively refines the model's reasoning to genuinely attend to audio properties.

This iterative process is motivated by the emergent audio Chain-of-Thought (CoT) capability observed after the cold-start phase. Our goal is to maintain and enhance this ability. We first construct a new set of perception-grounded questions based on our existing audio data. Then, at each iteration t, we use the model from the previous iteration ($\pi_{\theta_{t-1}}$) to perform self-distillation, generating new reasoning chains for this data.

Self-Distillation with Acoustic Reasoning. At each iteration t, we begin by curating audio data that strongly emphasizes perceptual analysis. Given an audio dataset $\mathcal{D}_{\text{audio}}$, we select examples (x_{audio}, q)

where answering question q requires direct acoustic feature analysis rather than high-level semantic understanding. This selection prioritizes tasks demanding attention to timbral qualities, temporal patterns, pitch contours, rhythmic structures, and other low-level auditory properties, ensuring the model cannot rely on textual surrogates. For each selected audio-question pair (x_{audio}, q), we prompt the current model π_{θ_t} to generate reasoning chains that explicitly reference acoustic features:

$$(r^{(i)}, a^{(i)}) \sim \pi_{\theta_t}(\cdot \mid x_{\text{audio}}, q), \quad i = 1, \dots, K \tag{4}$$

We sample K candidate responses and filter them using quality criteria that verify: (1) acoustic grounding—reasoning explicitly mentions perceptual features rather than textual descriptions; (2) logical coherence—reasoning steps follow sound inferential structure; and (3) answer correctness—final answers align with ground truth when available. This filtering yields a curated dataset $\mathcal{D}_t^{\text{audio-cot}}$ of acoustically-grounded reasoning chains.

Multimodal Supervised Refinement. We perform supervised fine-tuning on the distilled acoustic reasoning data, combined with original textual reasoning data to preserve existing capabilities:

$$\mathcal{L}_{\text{SFT}}^{(t)} = \mathbb{E}_{\mathcal{D}_t^{\text{audio-cot}}} \left[\log \pi_\theta(r, a \mid x_{\text{audio}}, q)\right] + \mathbb{E}_{\mathcal{D}_{\text{task}}} \left[\log \pi_\theta(r, a \mid q)\right] \tag{5}$$

This joint training anchors reasoning to acoustic properties while maintaining textual reasoning proficiency.

Multimodal Reinforcement Learning. We further refine the model through reinforcement learning on both audio and text tasks with carefully designed reward structures.

For text questions, we employ standard binary verification:

$$R_{\text{text}}(r, a) = \begin{cases} 1, & \text{if } a = a^* \\ 0, & \text{else} \end{cases} \tag{6}$$

For audio questions, we combine format and accuracy rewards:

$$R_{\text{audio}}(r, a) = 0.8 \times \begin{cases} 1, & \text{if } a = a^* \\ 0, & \text{else} \end{cases} + 0.2 \times \begin{cases} 1, & \text{if reasoning present in } r \\ 0, & \text{else} \end{cases} \tag{7}$$

This design prioritizes answer correctness (0.8 weight) while incentivizing reasoning generation (0.2 weight), preventing the model from reverting to direct responses. The combined optimization objective is:

$$\mathcal{L}_{\text{RLVR}}^{(t)} = \mathbb{E}_{\mathcal{D}_{\text{audio}}} \left[R_{\text{audio}}(r, a)\right] + \mathbb{E}_{\mathcal{D}_{\text{task}}} \left[R_{\text{text}}(r, a)\right] \tag{8}$$

Iterative Refinement. We repeat this cycle for T iterations, with each iteration t producing model $\pi_{\theta_{t+1}}$ that generates progressively more acoustically-grounded reasoning. As iterations advance, the model's reasoning chains shift from textual surrogates—such as inferring emotion from "lyrics mentioning sadness"—to genuine acoustic analysis—such as "minor key progressions and descending melodic contours." This iterative distillation progressively transforms the model's reasoning substrate from linguistic to acoustic grounding.

The final model π_{θ_T} achieves the desired capability: generating extended reasoning chains that genuinely attend to audio properties, thereby unlocking test-time compute scaling benefits in the audio domain.

4.3 Implement Details

RL Data Curation and Filtering Details. To construct the dataset for the RL phase, we extract text QA and audio data spanning diverse tasks and topics. We then filter these questions to identify a high-quality, challenging subset. Using the model from the $t-1$ iteration, we sample $k=8$ responses for each question ($pass@8$). A question is selected for the RL dataset if the number of correct passes falls within the range of $[3, 6]$. This filtering mechanism ensures we select for problems that are relatively difficult, filtering out both overly simple questions (where $pass@8 > 6$) and potentially harmful or nonsensical questions (where $pass@8 < 3$).

RL Implementation Details We employ an on-policy Proximal Policy Optimization framework [22] with binary verification rewards: responses receive a reward of 1.0 when matching verified solutions and 0.0 otherwise. Critically, we remove reference model KL penalties by setting the penalty coefficient to zero, allowing the model to freely explore reasoning strategies without being constrained by its initialization distribution. During training, we sample 16 candidate responses per prompt, assigning rewards exclusively at the final token position to encourage complete reasoning trajectories. We configure PPO with a clipping parameter of 0.2 and set both the discount factor and GAE lambda to 1.0, training on sequences up to 10,240 tokens to accommodate extended deliberation.

5 Evaluation

Having established that audio intelligence can indeed benefit from deliberate reasoning, we now present a comprehensive empirical evaluation of Step-Audio-R1. Our assessment rigorously examines its capabilities across a spectrum of complex audio tasks, structured into two key benchmarks: the Evaluation on Speech-to-Text Benchmarks, which measures understanding and reasoning from acoustic signals, and the Evaluation on Speech-to-Speech Benchmarks, which assesses the model's ability to perform generative and interactive reasoning in real-time spoken dialogue scenarios within the auditory domain.

5.1 Evaluation on Speech-to-Text Benchmarks

Table 1: Performance comparison (in %) on speech-to-text benchmarks across Big Bench Audio, Spoken MQA, MMSU, MMAU, Wild Speech, and Average Score.

Model	**Avg.**	Big Bench Audio	Spoken MQA	MMSU	MMAU	Wild Speech
Step-Audio 2	68.3	59.1	88.8	64.3	78.0	51.1
Gemini 2.5 Pro	81.5	96.1	94.8	79.3	77.4	60.0
Gemini 3 Pro	**85.1**	92.1	**95.3**	**82.9**	**78.9**	**76.4**
Step-Audio-R1	83.6	**98.7**	95.2	75.9	77.7	70.6

This section evaluates the speech understanding and reasoning capabilities of Step-Audio-R1 against several state-of-the-art baselines: the powerful large-language model Gemini 2.5 Pro, the newly released Gemini 3 Pro, our own previous-generation model Step-Audio 2, and the base Step-Audio-R1 model. The assessment is conducted across a comprehensive suite of benchmarks designed to probe advanced audio intelligence. These include MMSU [23] and MMAU [24] for expert-level audio understanding and reasoning, Big Bench Audio[1] for complex multi-step logical reasoning from audio, Spoken MQA [25] for mathematical reasoning with verbally expressed problems, and Wild Speech [26] for evaluating conversational speech.

As shown in Table 1, Step-Audio-R1 achieves an average score of 83.6%, significantly outperforming Gemini 2.5 Pro while being slightly lower than Gemini 3 Pro. This competitive performance confirms that our MGRD approach effectively enhances deep audio comprehension.

5.2 Evaluation on Speech-to-Speech Benchmarks

Table 2: Performance comparison of representative models on the Big Bench Audio speech-to-speech benchmark. The benchmark comprises two evaluation metrics: the Speech Reasoning Performance Score (%), measuring the model's reasoning ability over spoken content, and the first-packet Latency (seconds) metric, quantifying response speed as an indicator of dialogue fluency.

Model	Speech Reasoning Performance Score (%)	Latency (seconds)
GPT-4o mini Realtime	69	0.81
GPT Realtime 0825	83	0.98
Gemini 2.5 Flash Live	74	0.64
Gemini 2.5 Flash Native Audio Dialog	92	**0.63**
Step-Audio-R1 Realtime	**96.1**	0.92

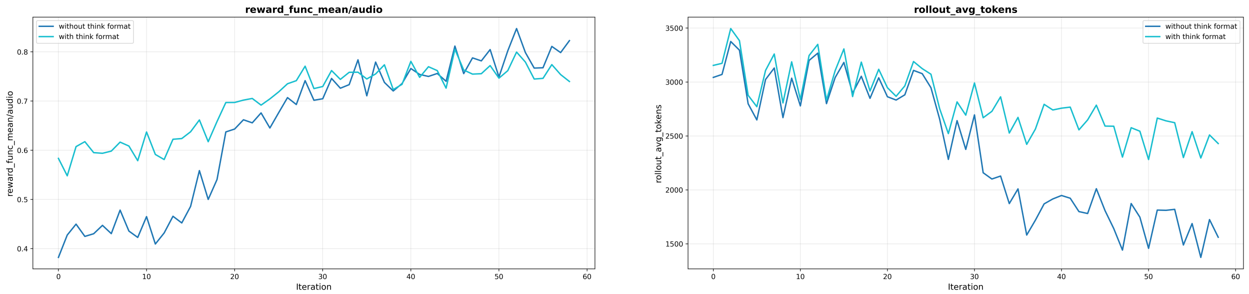
In this section, we evaluate the model's performance on the Big Bench Audio speech-to-speech benchmark. This benchmark consists of two major dimensions, namely the speech reasoning performance score, which assesses the model's ability to perform reasoning over spoken content, and the latency metric, which measures response speed as an indicator of dialogue fluency. Following the design of the listen-while-thinking[27] and think-while-speaking[28] architecture, we adapt Step-Audio-R1 into Step-Audio-R1 Realtime, attaining high-quality reasoning together with rapid responsiveness. According to Table 2, Step-Audio-R1 Realtime reaches a speech reasoning performance score of 96.1%, outperforming exemplary closed-source systems including GPT Realtime 0825 and Gemini 2.5 Flash Native Audio Dialog. Besides, Step-Audio-R1 Realtime achieves a first-packet latency of 0.92 s, maintaining sub-second responsiveness that represents a highly competitive interaction performance among contemporary audio language models. These results demonstrate that Step-Audio-R1 Realtime integrates real-time responsiveness with advanced reasoning capabilities, highlighting its potential for building efficient, intelligent, and interactive large audio language models.

6 Empirical Study and Analysis

6.1 Extended Reasoning Benefits Audio: Evidence from Format Reward Ablation

To validate the necessity of our composite reward design for audio tasks, we conduct an ablation study comparing training with and without the format reward component. The results reveal crucial insights into how reward structure shapes model behavior in audio reasoning tasks.

[1] `https://huggingface.co/datasets/ArtificialAnalysis/big_bench_audio`



(a) Mean reward evolution

(b) Reasoning token count

Figure 4: Impact of format rewards on audio reasoning training. (a) Format rewards enable faster and more stable convergence to high reward values. (b) Without format rewards, models exhibit systematic reasoning collapse, reducing generated tokens from 3000 to below 1500.

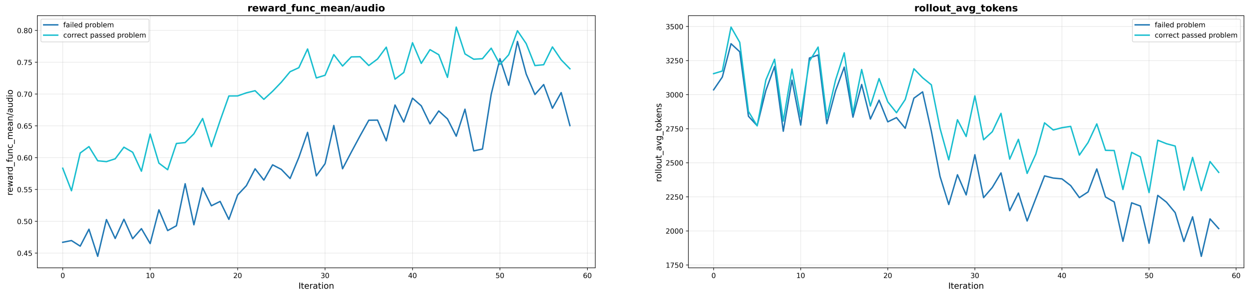
Format Reward Drives Stable Convergence. Figure 4a presents the evolution of mean reward on audio tasks across training iterations. Both configurations eventually converge to similar reward levels (approximately 0.75-0.80), but their trajectories differ significantly. The model *with* think format reward (cyan line) achieves stable performance earlier, reaching the 0.70 threshold around iteration 35-40, while the model *without* format reward (blue line) requires nearly 25 iterations to reach comparable performance. More critically, the format-rewarded model maintains more stable training dynamics in later iterations (30-60), whereas the baseline exhibits higher variance and occasional performance drops. This stability advantage translates to meaningful performance gains: on the MMAU benchmark, incorporating the format reward improves accuracy from 76.5 to 77.7.

Format Rewards Prevent Reasoning Collapse. Figure 4b reveals a striking phenomenon that explains the performance difference. Without format rewards, the model exhibits systematic collapse of reasoning length: starting from approximately 3000 tokens in early iterations, it progressively decays to below 1500 tokens by iteration 60, with a particularly sharp decline after iteration 30. In contrast, the model with format rewards maintains substantially longer and more stable reasoning chains throughout training, consistently generating 2300-2800 tokens even in later iterations. This 50-80% increase in reasoning length is not merely superficial verbosity—the accompanying performance improvements on MMAU confirm that these extended deliberations contain meaningful acoustic analysis.

The collapse pattern reveals a critical failure mode: without explicit incentives for reasoning generation, reinforcement learning naturally gravitates toward the most token-efficient strategy—direct answers without deliberation. This optimization pressure directly contradicts our goal of developing genuine reasoning capabilities. The think format reward component acts as a crucial regularizer, ensuring the model maintains extended thought chains even when pure accuracy-based rewards might prefer brevity.

Extended Reasoning Improves Audio Understanding. Most importantly, these training dynamics yield a fundamental shift in model capabilities: Step-Audio-R1 with extended reasoning chains consistently outperforms variants with minimal or no deliberation. This validates the central thesis of our work—that audio intelligence *can* benefit from extended deliberation when reasoning is properly grounded in acoustic properties. The performance gap between models with full reasoning (MMAU: 77.7) versus abbreviated or absent reasoning demonstrates that test-time compute scaling, once considered incompatible with audio tasks, now provides measurable advantages. This breakthrough confirms that the historical performance degradation with reasoning in audio models stemmed not from fundamental incompatibility, but from inadequate grounding mechanisms—a problem our MGRD framework successfully addresses.

6.2 Strategic Data Selection: Quality Over Quantity in Audio RL



(a) Mean reward evolution

(b) Reasoning token count

Figure 5: Impact of data selection strategies on audio reasoning training. (a) Training on moderately difficult problems (correct passed) achieves higher and more stable rewards compared to failed problems, which collapse after iteration 50. (b) Moderately difficult problems sustain reasoning generation (2300-2800 tokens) , while failed problems show a progressive decline, settling around 1800-2000 tokens by iteration 60

We discover that careful data curation proves more critical than dataset volume for audio reasoning tasks. Through comparing three data selection strategies, we reveal what constitutes effective training data for MGRD.

Comparing Selection Criteria. We evaluate two distinct data selection approaches for the RL phase: (1) *Consistently-failed problems*: questions where the SFT model from iteration $t-1$ fails all $k = 8$ sampled attempts ($pass@8 = 0$); (2) *Moderately difficult problems*: questions where correct passes fall within $[3, 6]$ out of 8 attempts, as described in our MGRD framework. Additionally, we experiment with (3) *Unfiltered scaling*: expanding the audio RL dataset to 200K examples without difficulty-based selection.

Learning from Partial Success Outperforms Learning from Failure. Figure 5a reveals striking differences in training dynamics. Models trained on moderately difficult problems (cyan line, "correct passed problem") achieve substantially higher and more stable rewards, converging to approximately 0.75-0.80 by iteration 20 and maintaining this performance throughout training. In sharp contrast, models trained exclusively on consistently-failed problems (blue line, "failed problem") exhibit significantly lower rewards (0.45-0.70) with higher variance and unstable convergence, eventually collapsing after iteration 50.

This performance gap stems from fundamental differences in learning signals. Problems where the SFT model consistently fails often indicate inherent ambiguity or insufficient information in the audio modality—for instance, inferring a car's brand from engine sounds alone, a task trivial in vision but nearly impossible from audio. Without correct reasoning exemplars in sampled trajectories, the model explores blindly, unable to distinguish between genuine acoustic limitations and solvable challenges. Moderately difficult problems, conversely, provide a crucial mix: some responses demonstrate correct acoustic reasoning paths while others reveal common failure modes. This combination enables effective policy gradient updates—the model learns both successful reasoning strategies and mistakes to avoid, while naturally filtering out acoustically ambiguous questions.

Reasoning Complexity Reflects Learning Quality. Figure 5b corroborates this finding through reasoning length analysis. Initially, both strategies generate similar reasoning lengths (approximately 3000-3500 tokens in early iterations), as they start from the same SFT checkpoint. However, their trajectories diverge significantly after iteration 20. Models trained on moderately difficult problems (cyan line) maintain substantial reasoning chains, stabilizing at 2300-2800 tokens throughout later training, demonstrating sustained deliberation. Models trained on consistently-failed problems (blue line), however, show progressive decline: reasoning length gradually decreases from iteration 20 onward, eventually settling around 1800-2000 tokens by iteration 60—a 30-40% reduction from the moderately difficult setting.

This divergence reveals how data quality shapes reasoning behavior over time. While both models initially maintain extended thinking inherited from SFT, continued training on consistently-failed problems gradually erodes this capability. The absence of successful reasoning exemplars provides no positive reinforcement for extended deliberation, causing the model to progressively abandon lengthy reasoning chains. In contrast, moderately difficult problems—which contain both successful and failed attempts—sustain the model's reasoning complexity by rewarding extended acoustic analysis that leads to correct answers.

Scale Without Strategy Provides No Benefit. Most surprisingly, we experiment with scaling the audio RL dataset to 200K examples—over 10× our curated subset—and observe no performance improvement. This null result carries important implications: in audio reasoning tasks, data quality substantially outweighs quantity. Indiscriminate scaling introduces noise from acoustically ambiguous or inherently unsolvable problems, diluting the learning signal from genuinely informative examples. The effectiveness of challenging-but-solvable problems suggests that successful audio reasoning requires careful curriculum design rather than brute-force data scaling.

6.3 Self-Cognition Correction Through Iterative Refinement

A critical challenge emerges when training Audio LLMs on massive textual data: models tend to develop incorrect self-cognition [12, 13]. Due to the dominance of text-only patterns in the training corpora, these models frequently claim inability to process audio inputs by stating "I cannot hear sounds" or "I am a text model." This misalignment between actual capabilities and self-perception severely undermines user experience and model utility. We address this systematic bias through a multi-stage correction pipeline combining iterative self-distillation with preference optimization.

Iterative Self-Distillation with Cognition Filtering. Our correction process begins with targeted data curation. We construct a dataset of audio perception queries specifically designed to elicit self-cognition responses—questions about sound identification, audio quality assessment, and acoustic property analysis. During the self-distillation SFT iterations, we employ an LLM judge to filter responses exhibiting incorrect self-cognition. The judge evaluates whether the model appropriately acknowledges its audio processing capabilities or incorrectly identifies as text-only. Only responses with correct self-cognition pass to the next training round, progressively reinforcing accurate self-perception while eliminating erroneous beliefs.

Preference Optimization for Final Correction. Following the filtered self-distillation phase, we apply DPO [29] for precise calibration. We construct 8,000 preference pairs through self-distillation: positive examples comprise responses where the model correctly acknowledges and utilizes its audio capabilities, while negative examples contain responses claiming text-only limitations. This contrastive learning directly targets the remaining self-cognition errors, teaching the model to consistently choose responses aligned with its true multimodal nature. Despite the relatively modest dataset size, this targeted approach proves remarkably effective due to the specificity of the correction task.

Table 3: Self-cognition error rates across training stages on our constructed test set of 5,000 diverse audio perception samples. Error rate measures the percentage of responses where the model incorrectly claims inability to process audio.

Training Stage	Self-Cognition Error Rate
Base model	6.76%
Iterative Self-Distillation	2.63%
Iterative Self-Distillation + DPO	0.02%

Progressive Error Reduction. Table 3 demonstrates the effectiveness of our multi-stage approach. The base model exhibits noticeable self-cognition errors (6.76%), reflecting the bias from text training data. Through iterative self-distillation, we successfully reduce the error rate to 2.63% by filtering misaligned responses and reinforcing correct self-perception. However, the most dramatic improvement comes from the final DPO alignment with 8,000 preference pairs: error rates plummet to near-zero (0.02%), effectively eliminating self-cognition misalignment. This final stage proves crucial—while self-distillation significantly improves cognition, only explicit preference optimization achieves complete correction. The efficiency of this approach highlights the power of targeted preference learning for addressing specific behavioral biases.

The success of this approach reveals an important insight: self-cognition errors are not fundamental model limitations but learned biases from training data distribution. Through systematic correction combining iterative refinement with targeted preference optimization, we demonstrate that models can maintain accurate self-perception even when pretrained on predominantly text data. This correction is essential for Step-Audio-R1's deployment—users expect the model to confidently engage with audio inputs rather than apologetically claim incapability.

7 Conclusion

In this work, we address the challenging problem where audio language models historically fail to benefit from long reasoning processes, often performing worse as the reasoning length increases. We identify the primary cause of this failure as "textual surrogate reasoning"—a persistent tendency for models to reason based on text descriptions, such as transcripts or captions, rather than focusing on actual acoustic properties. We introduce Step-Audio-R1, the first model to successfully unlock and benefit from deliberate thinking in the audio domain. Our core contribution is Modality-Grounded Reasoning Distillation (MGRD), an iterative framework that progressively shifts the model's reasoning basis from text-based patterns to genuine acoustic analysis. Comprehensive evaluations confirm that Step-Audio-R1 outperforms strong baselines, including Gemini 2.5 Pro, and achieves performance comparable to the state-of-the-art Gemini 3 Pro across a wide range of complex audio understanding and reasoning benchmarks. These results provide clear evidence that reasoning is a capability that works across modalities; when properly connected to the correct input, extended reasoning transforms from a weakness into a powerful asset for audio intelligence, opening new paths for building truly multimodal systems.

8 Contributors

Core Contributors: Fei Tian[1,*,†], **Xiangyu Tony Zhang**[1,3], **Yuxin Zhang**[1,4], **Haoyang Zhang**[1,2], **Yuxin Li**[1,2], **Daijiao Liu**[1,3]

Contributors: Yayue Deng[1], **Donghang Wu**[1,2], **Jun Chen**[1], **Liang Zhao**[1], **Chengyuan Yao**[1], **Hexin Liu**[2], **Eng Siong Chng**[2], **Xuerui Yang**[1], **Xiangyu Zhang**[1], **Daxin Jiang**[1], **Gang Yu**[1]

[1]StepFun [2]Nanyang Technological University [3]University of New South Wales
[4]Shanghai Jiao Tong University
[*]Corresponding authors: tianfei@stepfun.com [†]Project Leader

References

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

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

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

[4] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[5] Bin Wang, Bojun Wang, Changyi Wan, Guanzhe Huang, Hanpeng Hu, Haonan Jia, Hao Nie, Mingliang Li, Nuo Chen, Siyu Chen, et al. Step-3 is large yet affordable: Model-system co-design for cost-effective decoding. *arXiv preprint arXiv:2507.19427*, 2025.

[6] Wei Shen, Jiangbo Pei, Yi Peng, Xuchen Song, Yang Liu, Jian Peng, Haofeng Sun, Yunzhuo Hao, Peiyu Wang, Jianhao Zhang, et al. Skywork-r1v3 technical report. *arXiv preprint arXiv:2507.06167*, 2025.

[7] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

[8] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

[10] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in

language models. In *The Eleventh International Conference on Learning Representations*.

[11] Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025.

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

[13] Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

[14] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[15] Jiajun Fan, Roger Ren, Jingyuan Li, Rahul Pandey, Prashanth Gurunath Shivakumar, Ivan Bulyko, Ankur Gandhe, Ge Liu, and Yile Gu. Incentivizing consistent, effective and scalable reasoning capability in audio llms via reasoning process rewards. *arXiv preprint arXiv:2510.20867*, 2025.

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

[17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL `https://arxiv.org/abs/2201.11903`.

[18] Google DeepMind. Gemini 3. `https://deepmind.google/models/gemini/`, 2025. Accessed: 2025.

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

[20] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

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

[22] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[23] Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*, 2025.

[24] S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

[25] Chengwei Wei, Bin Wang, Jung-jae Kim, and Nancy F Chen. Towards spoken mathematical reasoning: Benchmarking speech-based models over multi-faceted math problems. *arXiv preprint arXiv:2505.15000*, 2025.

[26] Jian Zhang, Linhao Zhang, Bokai Lei, Chuhan Wu, Wei Jia, and Xiao Zhou. Wildspeech-bench: Benchmarking audio llms in natural speech conversation. *arXiv preprint arXiv:2506.21875*, 2025.

[27] Donghang Wu, Haoyang Zhang, Chen Chen, Tianyu Zhang, Fei Tian, Xuerui Yang, Gang Yu, Hexin Liu, Nana Hou, Yuchen Hu, and Eng Siong Chng. Chronological thinking in full-duplex spoken dialogue language models, 2025. URL `https://arxiv.org/abs/2510.05150`.

[28] Donghang Wu, Haoyang Zhang, Jun Chen, Hexin Liu, Eng Siong Chng, Fei Tian, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, Gang Yu, et al. Mind-paced speaking: A dual-brain approach to real-time reasoning in spoken language models. *arXiv preprint arXiv:2510.09592*, 2025.

[29] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL `https://arxiv.org/abs/2305.18290`.