

Audio-Reward-R1: Reasoning Reward Model for Unified Audio Understanding and Generation

Hao Gu¹, Xiaocheng Yang^{1,†}, Jinyang Wu³, Hongbo Zhao²,
Yunzhe Hao¹, Songcheng Xu¹, Ye Bai¹, Zhao You¹, Shidong Shang¹,

¹Tencent Inc., ²University of Chinese Academy of Sciences,

³Department of Automation, Tsinghua University,
orangexyang@tencent.com

Abstract

Multimodal Reward Models are critical for aligning Multimodal Large Language Models with human preferences. However, existing reward models offer limited interpretability and struggle to deliver reliable preference judgments in audio-centric scenarios. To address this gap, we propose **Audio-Reward-R1**, a reasoning-based reward model that supports both audio understanding and audio generation preference tasks. Audio-Reward-R1 formulates reward modeling as a structured reasoning process, producing dimension-level analyses prior to the final preference judgment. To enable reliable reasoning supervision, we construct a high-quality reasoning-enhanced audio preference dataset with structured reasoning traces aligned with preference annotations. Audio-Reward-R1 is first trained via supervised fine-tuning as a cold-start on this reasoning-enhanced dataset, followed by reinforcement learning with verifiable rewards. Finally, we introduce **Audio-Reward-Bench**, a human-verified evaluation benchmark for evaluating audio reward models across both audio understanding and audio generation preference tasks. Extensive experimental results demonstrate the effectiveness and reliability of Audio-Reward-R1. Code is available at https://github.com/ucas-hao/audio_reward_r1.

1 Introduction

Multimodal reward models (Wang et al., 2024, 2025c) play a critical role in aligning multimodal Large Language Models with human preferences, providing crucial reward signals to guide model training (Ouyang et al., 2022; Ji et al., 2024; Rafailov et al., 2023) and inference (Wang et al., 2025b). While reward modeling has advanced rapidly in text (Chen et al., 2025; Guo et al., 2025b) and visual (Zhang et al., 2025c; Wang et al., 2025c) domains, progress in the audio modality remains

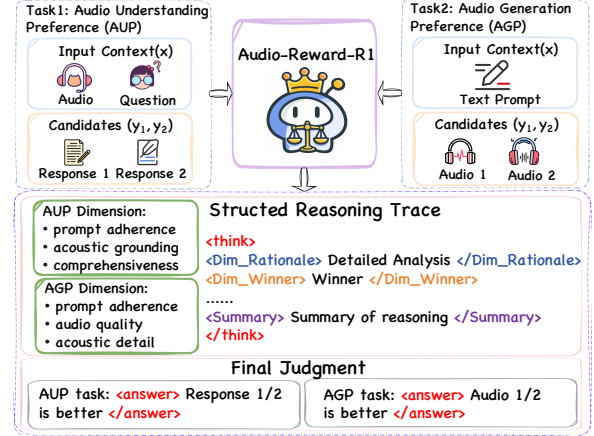


Figure 1: **Overview of Audio-Reward-R1.** Our model generates structured reasoning with dimension-level analysis for unified preference modeling across audio understanding and generation tasks.

comparatively limited (Zhang et al., 2025b).

Early efforts in audio reward modeling primarily rely on carefully engineered prompts that induce models to output scalar scores or rankings for candidate outputs (Manakul et al., 2025), or on supervised fine-tuning schemes that train models to predict explicit quality scales (Kuan et al., 2026; Wang et al., 2025a). While direct and often effective, these approaches lack interpretability, offering no intermediate reasoning steps to justify the model’s decisions. More recently, reasoning-based reward modeling (Chen et al., 2025; Ji et al., 2025) has gained increasing attention, incorporating explicit reasoning into reward modeling, enabling models to generate intermediate reasoning traces before producing final judgments. For example, SpeechJudge (Zhang et al., 2025b) frames speech naturalness assessment as a reasoning task and leverages reinforcement learning to align model judgments with human preferences. However, existing methods remain limited. *First*, they are typi-

[†]Corresponding author.

cally restricted to narrow settings, such as pairwise comparison of speech naturalness or audio quality, and do not generalize to broader audio understanding and generation preference scenarios. *Second*, the construction of reasoning traces often relies on directly providing ground-truth preferences to a teacher model, which risks encouraging post-hoc rationalizations rather than faithful reasoning, and does not ensure consistency between intermediate reasoning and final judgments.

In this paper, we address these limitations by proposing **Audio-Reward-R1**, a unified reasoning-based audio reward model that supports both Audio Understanding Preference (AUP) and Audio Generation Preference (AGP). In AUP, the reward model evaluates competing textual responses conditioned on an audio input by analyzing acoustic evidence and semantic alignment; in AGP, it compares audio samples generated from the same prompt with respect to perceptual quality and acoustic details. An overview of Audio-Reward-R1 is illustrated in Figure 1. To enable reliable reasoning supervision, we construct a high-quality reasoning-augmented audio preference dataset by generating reasoning traces with a teacher model, refining traces with minor dimension-level mismatches, and retaining only those fully aligned with ground-truth annotations at both dimension and overall levels. Audio-Reward-R1 is first fine-tuned on this reasoning-augmented dataset as a cold-start, followed by reinforcement learning with verifiable rewards, resulting in a model that produces structured, dimension-level reasoning for preference judgment. We further introduce **Audio-Reward-Bench**, a human-verified evaluation benchmark for audio reward modeling covering both audio understanding and audio generation preference scenarios.

In summary, our main contributions are:

- We construct a reasoning-enhanced dataset for audio reward modeling that jointly supports audio understanding and audio generation preference, together with a carefully curated evaluation benchmark, **Audio-Reward-Bench**.
- We propose **Audio-Reward-R1**, the first unified reasoning-based audio reward model that explicitly decomposes preference judgments into structured, dimension-level analyses.
- Extensive experiments and ablation studies demonstrate the effectiveness and superiority of

our approach across audio understanding and audio generation preference tasks.

2 Related work

2.1 Multimodal Reward Model

Reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Ji et al., 2024; Yu et al., 2025; Wu et al., 2026c; Yang et al., 2026; Wu et al., 2026b; Feng et al., 2026a) has emerged as an effective paradigm for aligning MLLMs with human preferences. Prior studies have demonstrated that RLHF can substantially improve multimodal understanding (Feng et al., 2026b, 2025) and reasoning capabilities (Li et al., 2025; Huang et al., 2025), while also enhancing output quality and controllability in multimodal generation tasks. A key component in the alignment process is the reward model, which serves to model human preference and provide feedback signals that steer the model towards more desirable behaviors. Recently, an increasing number of reward models have been proposed for multimodal tasks. For text-to-image generation, ImageReward (Xu et al., 2023) evaluates semantic alignment and visual quality. UnifiedReward (Wang et al., 2025c) extends reward modeling to both visual understanding and generation. In audio domain, SpeechJudge (Zhang et al., 2025b) focuses on speech naturalness evaluation and employs reinforcement learning to align model judgments with human preferences. However, existing approaches in the audio domain are typically restricted to narrow evaluation settings, such as speech naturalness. In contrast, our work introduces the first unified reward model that jointly supports both audio understanding and audio generation preference scenarios.

2.2 Reward Model Training

There are generally three main approaches to reward modeling (Jin et al., 2025; Wu et al., 2026a). The first approach relies on prompting a language model or a multimodal models to directly output preference rankings or scores for candidate outputs by carefully designing prompts (Xiong et al., 2025). However, this method (Zhang et al., 2025c) heavily depends on the model’s instruction-following and comprehension ability. The second approach (Jin et al., 2025) attaches a lightweight reward head to the latent representation of a language model and trains it to directly output a scalar score. Although computationally efficient, this approach discards

the intermediate reasoning process and thus lacks interpretability (Liu et al., 2024; Lou et al., 2024). More recently, a third line of work has begun to incorporate explicit reasoning into reward modeling, encouraging the reward model to generate intermediate reasoning traces before producing a final preference judgment. Some studies, exemplified by RM-R1 (Chen et al., 2025) and R1-Reward (Zhang et al., 2025c), introduce reinforcement learning techniques such as GRPO (Guo et al., 2025a) to optimize structured reasoning with verifiable rewards. In this work, we extend reasoning-based reward modeling to the audio domain and propose Audio-Reward-R1, supporting both audio understanding and audio generation preference scenarios.

3 Dataset Construction

3.1 Task Definition

We consider reasoning-based reward modeling in a pairwise comparison setting. Each instance is represented as (x, y_1, y_2, r, l) , where x denotes the input context, (y_1, y_2) are two candidate outputs, r is a structured reasoning trace that compares the two candidates, and $l \in \{y_1, y_2\}$ denotes the final judgment indicating the preferred candidate. The reward model learns to generate (r, l) conditioned on (x, y_1, y_2) . Based on this unified formulation, we focus on two audio-centric tasks.

Audio Understanding Preference (AUP) This task evaluates the model’s ability to assess answer quality grounded in audio understanding. The input x consists of an audio clip paired with a question about its content, and the candidates (y_1, y_2) are textual answers to the same question. The reward model reasons over the audio evidence, compares the two answers, and generates a reasoning trace followed by a final preference judgment.

Audio Generation Preference (AGP) This task evaluates generated audio conditioned on a text prompt. The input x is a prompt, and the candidates (y_1, y_2) are two audio outputs generated from it. The reward model compares the samples with respect to prompt alignment and perceptual quality, generates a reasoning trace, and outputs a final preference decision.

3.2 Data Construction Pipeline

Step 1: Data Collection Our dataset is built upon Align-Anything (Ji et al., 2024), which provides large-scale multimodal preference annota-

tions. Each instance contains an input context, two candidate outputs, fine-grained 1–5 ratings for each candidate across multiple evaluation dimensions, and an overall pairwise preference.

While Align-Anything annotates a wide range of dimensions for cross-modal consistency, only a subset is relevant to audio-centric reasoning. We retain dimensions most critical to audio understanding and generation. For Audio Understanding Preference (AUP), we focus on *prompt adherence*, *acoustic grounding*, and *comprehensiveness*. For Audio Generation Preference (AGP), we retain *prompt adherence*, *audio quality*, and *acoustic detail*. Together, these dimensions provide a compact and expressive basis for audio reward modeling.

Step 2: Quality Filtering and Benchmark Construction

To improve annotation reliability, we apply an automatic quality auditing step using Gemini-3-pro-preview to remove low-quality samples. The auditing criteria and prompt templates are described in Appendix A.1. From the filtered data, we construct a held-out evaluation benchmark, termed **Audio-Reward-Bench**, consisting of 2,000 AUP instances and 1,000 AGP instances. The benchmark is **balanced** with respect to the two preference outcomes and each instance is manually verified to ensure evaluation reliability. Details of benchmark construction are provided in Appendix A.2. All remaining samples are used for training and form the corpus $\mathcal{D}_{\text{Train}}$.

Step 3: Reasoning Response Generation

To support reasoning-based reward modeling, we construct structured reasoning traces that explicitly justify preference judgments through dimension-level analyses. For each instance, the reasoning process decomposes the comparison into a fixed set of evaluation dimensions, produces a comparative rationale and a judgment for each dimension, and then aggregates these analysis into a final overall preference. Dimension-level judgments allow three outcomes—candidate 1 preferred, candidate 2 preferred, or tie—while the final decision is constrained to a binary choice. The overall construction pipeline is illustrated in Fig. 2(a).

We employ Gemini-3-pro-preview as the teacher model. For each instance in $\mathcal{D}_{\text{Train}}$, we query the teacher and retain reasoning traces whose overall preference and all dimension-level judgments exactly match the ground-truth annotations as reliable reasoning supervision. The rollout prompt templates are provided in Appendix A.3.

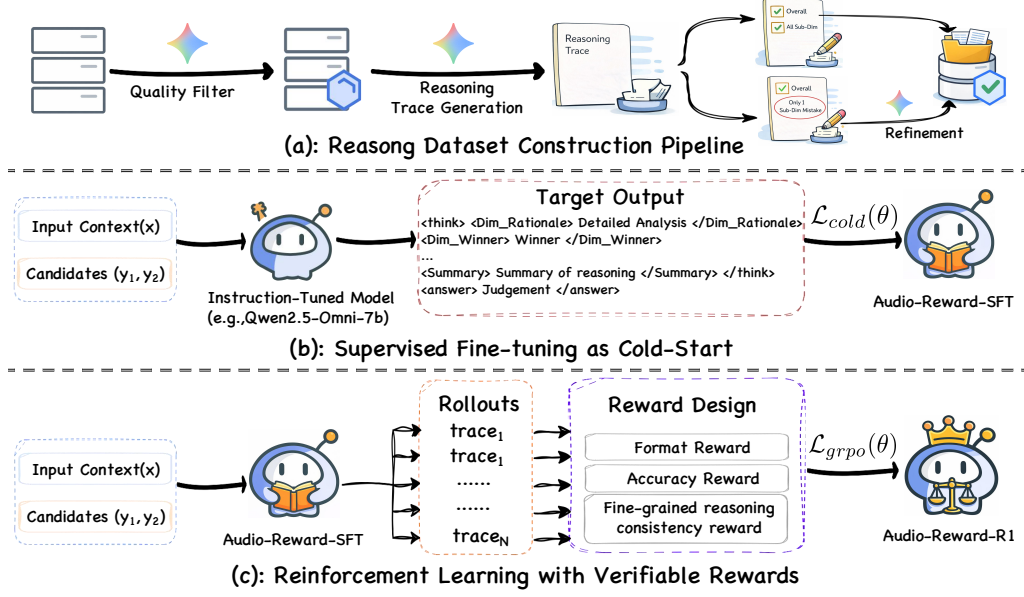


Figure 2: **Method Overview.** The framework consists of three components: (a) reasoning dataset construction; (b) supervised fine-tuning as cold-start; and (c) reinforcement learning with verifiable rewards.

In addition, we identify traces whose overall preference is correct but contain exactly one dimension-level mismatch with the ground-truth annotations. These cases reflect localized reasoning errors rather than fundamentally incorrect judgments. Instead of discarding them, we apply a refinement procedure: the teacher is given its previous reasoning output along with the ground-truth label for the incorrect dimension and instructed to revise accordingly. The refined traces, together with the fully consistent ones, form the high-quality reasoning dataset $\mathcal{D}_{\text{SFT}}$. The refinement prompts are provided in Appendix A.3.

Further discussion of our reasoning trace reliability is provided in Appendix A.5.

Step 4: Dataset Partitioning Based on the generated reasoning traces, we partition $\mathcal{D}_{\text{Train}}$ into two subsets. The $\mathcal{D}_{\text{SFT}}$ consists of balanced samples whose reasoning traces are fully aligned at both the dimension and overall levels. The remaining training samples form a complementary dataset, denoted as $\mathcal{D}_{\text{Comp}}$. Together with the held-out Audio-Reward-Bench, these splits constitute the final dataset used in our experiments. Dataset statistics are summarized in Appendix A.4.

4 Method

4.1 Overview

We present Audio-Reward-R1, a unified audio reward model that produces reliable preference judg-

ments with structured and interpretable reasoning. As illustrated in Fig.2(b)–(c), Audio-Reward-R1 is trained using a two-stage framework consisting of supervised fine-tuning as a cold start (Sec.4.2) followed by reinforcement learning with verifiable rewards (Sec. 4.3). The cold-start stage initializes the model with structured, dimension-level reasoning, while the subsequent reinforcement learning stage further strengthens its reasoning ability.

4.2 Supervised Fine-tuning as Cold-Start

For an instruction-tuned model (e.g., Qwen-2.5 Omni), it is quite intuitive to adapt it into a generative reward model through prompting. However, due to the absence of exposure to reward modeling during pre-training, these models often struggle to conduct consistent judgments.

To address this limitation, we introduce a supervised fine-tuning (SFT) stage as a cold-start. This stage trains the model to generate structured, dimension-level reasoning with a final preference decision, equipping it with the behavior required for reward modeling.

Specifically, we leverage the reasoning-augmented preference dataset $\mathcal{D}_{\text{SFT}}$ constructed in Sec. 3.2. Formally, given an input tuple (x, y_1, y_2) , where x denotes the input context and (y_1, y_2) are candidates, the model is trained to generate an output sequence:

$$\mathbf{z} = (z_1, z_2, \dots, z_T) = \text{Concat}(r, l) \quad (1)$$

where r represents the reasoning trace and l denotes the final preference judgment.

The training objective is defined as:

$$\mathcal{L}_{\text{cold}}(\theta) = - \sum_{t=1}^T \log p_{\theta}(z_t \mid x, y_1, y_2, z_{<t}), \quad (2)$$

where θ denotes the model parameters. This cold-start stage provides a reliable initialization for subsequent reinforcement learning.

4.3 Reinforcement Learning with Verifiable Rewards

4.3.1 Verifiable reward

During this stage, verifiable rewards play a crucial role in guiding the learning of ALLMs by providing explicit, rule-based feedback.

Format reward We introduce a format reward R_{fmt} to encourage structured and well-organized outputs, which is crucial for clarity, interpretability, and training stability. Although different tasks involve distinct evaluation dimensions, the output adheres to a unified format while instantiating task-specific evaluation dimensions. The model is required to generate responses that separate dimension-level rationales, corresponding judgments, a reasoning summary, and a final preference decision. The format reward is set to 1 if the output strictly adheres to the required structure with components correctly ordered, and 0 otherwise.

Accuracy reward The accuracy reward evaluates whether the model produces a correct final decision. Specifically, it checks whether the final judgment, enclosed within the <answer> tag matches the ground-truth preference (e.g., Audio/Response X is better). The accuracy reward R_{acc} is set to 1 if the judgment matches the ground truth and 0 otherwise.

Fine-grained reasoning consistency reward In addition to the format and accuracy rewards, we introduce a fine-grained consistency reward R_{fine} to supervise the model’s intermediate reasoning.

Using the audio generation preference task as an example, given a text prompt and two candidate audio clips, the model produces dimension-level preference predictions under three predefined criteria: prompt adherence, audio quality, and acoustic detail. The fine-grained reward compares each predicted winner with its ground-truth label, assigning

1 for each correctly predicted dimension and 0 otherwise. The total reward is obtained by summing the scores across all evaluated dimensions.

This reward encourages the model to produce reasoning that is not only overall correct but also consistent with the underlying evaluation criteria. By providing supervision at multiple granularities, the fine-grained consistency reward complements the accuracy reward and helps stabilize training by discouraging spurious reasoning patterns that lead to correct answers for incorrect reasons.

The overall verifiable reward R is formulated as:

$$R = R_{acc} + \alpha R_{fmt} + \beta R_{fine} \quad (3)$$

where α and β are weighting coefficients. In our experiments, both are set to 0.2 by default. The accuracy reward R_{acc} ensures correct final preference judgment, the format reward R_{fmt} enforces structured outputs, and the fine-grained consistency reward R_{fine} provides explicit supervision over intermediate reasoning, encouraging alignment between the reasoning process and the final judgment.

4.3.2 RL Training

Although SFT effectively adapts a general generative model into a reward model, it often overfits to fixed reasoning patterns and constrains the model’s ability to generalize reasoning for critical thinking (Chu et al., 2025), which is essential for reward modeling. To address this limitation, we adopt Group Relative Policy Optimization (GRPO) (Guo et al., 2025a) to further enhance reasoning capabilities. Given an input i , GRPO samples N candidate outputs $\{o^{(1)}, \dots, o^{(N)}\}$ from the previous policy $\pi_{\theta_{\text{old}}}$. Each output is evaluated using the verifiable rewards described in Sec. 4.3.1, yielding scalar scores $\{R^{(1)}, R^{(2)}, \dots, R^{(N)}\}$.

To measure the relative quality of each output within the sampled group, reward scores are standardized into advantages. Specifically, the advantage for the i -th output is computed as:

$$\hat{A}^{(i)} = \frac{R^{(i)} - \text{mean}(\{R^{(1)}, \dots, R^{(N)}\})}{\text{std}(\{R^{(1)}, \dots, R^{(N)}\})}. \quad (4)$$

GRPO then quantifies the magnitude of the policy update by comparing the likelihood of each sampled response under the new policy π_{θ} with that under the old policy $\pi_{\theta_{\text{old}}}$:

$$r^{(i)} = \frac{\pi_{\theta}(o^{(i)} \mid x)}{\pi_{\theta_{\text{old}}}(o^{(i)} \mid x)}. \quad (5)$$

To prevent excessively aggressive updates, the ratio is clipped to the interval $[1 - \delta, 1 + \delta]$. In addition, a KL divergence penalty is introduced to regularize the updated policy toward a fixed reference policy π_{ref} , with a weighting coefficient β .

The overall GRPO objective is defined as:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left[\min(r^i A_{i,t}, \text{clip}(r^i, 1 - \epsilon, 1 + \epsilon) A_{i,t}) - \beta D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}) \right] \quad (6)$$

Here $D_{\text{KL}}(\cdot \parallel \cdot)$ represents the KL divergence.

By combining normalized advantage estimation, clipped importance weighting, and reference-policy regularization, GRPO facilitates stable policy optimization and encourages the generation of high-quality, verifiably correct reasoning paths.

5 Experiments

5.1 Experiment Setup

Datasets. We use the dataset constructed in Sec. 3.2. The reasoning-enhanced dataset $\mathcal{D}_{\text{SFT}}$ is used for cold-start. The full training set $\mathcal{D}_{\text{Train}}$ is used for reinforcement learning. Audio-Reward-Bench serves as the held-out evaluation benchmark.

Baselines. We evaluate 14 models in a zero-shot manner in total, providing a diverse evaluation of model capabilities across different architectures. Open-source Large Audio Language Models includes Kimi-Audio-7B-Instruct (Ding et al., 2025), MiMo-Audio-7B-Instruct (Zhang et al., 2025a), Step-Audio-2-mini (Wu et al., 2025), Audio-Flamingo-3 (Goel et al., 2025), Qwen-audio-chat (Chu et al., 2023), Qwen2-Audio-7B-Instruct (Chu et al., 2024), Qwen2.5-Omni-7B (Xu et al., 2025a), Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b). Open-source Large Audio Reasoning Models includes Audio-Reasoner (Xie et al., 2025), MiMo-Audio-7B-Instruct-think (Zhang et al., 2025a), Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025b). The closed-source models include Gemini-2.5-flash, Gemini-2.5-pro and Gemini-3-pro-preview. More details are provided in Appendix B.

Training Details. We adopt Qwen2.5-Omni-7B (Xu et al., 2025a) as the base instruction-tuned model. We first perform supervised fine-tuning on $\mathcal{D}_{\text{SFT}}$ to obtain Audio-Reward-SFT. Then, we

further optimize Audio-Reward-SFT via GRPO on $\mathcal{D}_{\text{Train}}$, resulting in the final model, Audio-Reward-R1. For cold-start training, we use a batch size of 2, 2 gradient accumulation steps, a learning rate of 1×10^{-5} , and a warm-up ratio of 0.05. For GRPO training, we use a batch size of 1 with 8 gradient accumulation steps, a learning rate of 1×10^{-6} , a warm-up ratio of 0.01, and a KL penalty coefficient $\beta = 0.001$. The number of generated responses is set to $N = 8$. All experiments are conducted on 8 NVIDIA H20 GPUs.

Evaluation Metrics. Following SimpleVQA (Cheng et al., 2025) benchmark, we adopt rigorous metrics to assess the accuracy and reliability of the model’s predictions. Specifically, these metrics include: (1) **Accuracy (ACC)** measures whether the predicted answer matches the reference answer exactly without any contradiction. (2) **Not Attempted (NA)** identifies cases where the model does not attempt to answer, ensuring no contradictions. (3) **Incorrect (IN)** denotes instances where the predicted answer contradicts the reference answer. (4) **Correct Given Attempted (CGA)** computes the proportion of correct answers among those attempted by the model, reflecting its performance when engaged. (5) **F-score** provides a balanced evaluation that combines correctness and attempt success by computing the harmonic mean of "Accuracy" and "Correct Given Attempted".

5.2 Main Results

Table 1 presents the overall performance on Audio-Reward-Bench, revealing the following findings:

There is a clear performance gap between open-source and closed-source models. As shown in Table 1, the Gemini series consistently achieves ACC above 79% on the AUP task, demonstrating strong multimodal understanding capabilities. In contrast, most open-source baselines range between 60% and 70%. This gap suggests that open-source models still lag in generalizing to complex preference-based scenarios.

The AGP task is significantly more challenging than AUP. Most open-source models experience a sharp performance drop on AGP. We attribute this to the limited exposure to multi-audio processing scenarios during their training. As a result, these models struggle to compare preferences between two candidate audios and often produce refusals or invalid outputs. This limitation validates the necessity of our proposed Audio-Reward-R1.

Table 1: The overall performances on Audio-Reward-Bench across the Audio Understanding Preference (AUP) and Audio Generation Preference (AGP) tasks. All metrics are reported in percentage(%), and the best is in **bold**.

Models	Audio Understanding Preference (AUP)					Audio Generation Preference (AGP)				
	ACC↑	NA↓	IN↓	CGA↑	F-score↑	ACC↑	NA↓	IN↓	CGA↑	F-score↑
Open-source Audio Large Language Models										
Kimi-Audio-7B-Instruct	66.70	4.15	29.15	69.59	68.11	53.10	3.80	43.10	55.20	54.13
MiMo-Audio-7B-Instruct	67.65	0.10	32.25	67.72	67.68	54.20	0.40	45.40	54.42	54.31
Step-Audio-2-mini	64.85	7.90	27.25	70.41	67.52	55.90	5.60	38.50	59.22	57.51
Audio-Flamingo-3	65.60	12.55	21.85	75.01	69.99	45.20	0.90	53.90	45.61	45.40
Qwen-audio-chat	58.20	15.40	26.40	68.79	63.06	43.40	10.20	46.40	48.33	45.73
Qwen2-Audio-7B-Instruct	61.70	1.10	37.20	62.39	62.04	52.90	12.4	34.70	60.39	56.40
Qwen2.5-Omni-7B	69.80	0.95	29.25	70.47	70.13	57.50	5.30	37.20	60.72	59.07
Qwen3-Omni-30B-A3B-Instruct	74.70	0.20	25.10	74.85	74.77	59.80	0.20	40.00	59.92	59.86
Open-source Audio Large Reasoning Models										
Audio-Reasoner	60.65	17.90	21.45	73.87	66.61	31.00	34.10	34.90	47.04	37.37
MiMo-Audio-7B-Instruct-think	73.80	0.20	26.00	73.95	73.87	57.70	1.60	40.70	58.64	58.17
Qwen3-Omni-30B-A3B-Thinking	78.25	9.10	12.65	86.08	81.98	63.40	1.10	35.50	64.11	63.75
Closed-source Audio Large Language Models										
Gemini-2.5-flash	79.20	0.00	20.80	79.20	79.20	70.30	0.00	29.70	70.30	70.30
Gemini-2.5-pro	84.85	0.00	15.15	84.85	84.85	75.80	0.00	24.20	75.80	75.80
Gemini-3-pro-preview	86.15	0.00	13.85	86.15	86.15	78.60	0.00	21.40	78.60	78.60
Our Methods										
Audio-Reward-SFT	83.25	0.05	16.70	83.29	83.27	72.90	0.00	27.10	72.90	72.90
Audio-Reward-R1	87.90	0.00	12.10	87.90	87.90	80.10	0.00	19.90	80.10	80.10

Audio-Reward-R1 achieves the best performances on both AUP and AGP. Specifically, it attains F-scores of **87.90%** on AUP and **80.10%** on AGP. Our method not only outperforms all open-source baselines by a large margin but also surpasses the strongest closed-source model, Gemini-3-pro-preview. These results demonstrate superior preference alignment of our method.

6 Ablations and Analysis

6.1 Unified Training Strategy

In this section, we investigate the impact of unifying AUP and AGP tasks into a single model. We aim to verify whether the unified training strategy yields performance gains.

Setup As presented in Table 2, we conduct an ablation study systematically comparing three different training strategies: (1) Unified: the model is jointly trained on both AUP and AGP tasks; (2) Only AUP: the model is trained exclusively on the AUP task; (3) Only AGP: the model is trained exclusively on the AGP task. All strategies are evaluated under both SFT and SFT+RL settings. We report the accuracy (%) across fine-grained dimensions as well as the overall accuracy (%) for both

AUP and AGP tasks.

Results Experimental results in Table 2 demonstrate that the unified training strategy consistently outperforms single-task baselines. For instance, under the SFT+RL setting, the Unified model achieves the highest overall accuracy of 87.90% on the AUP task, surpassing the Only AUP baseline (86.75%). A similar performance advantage is observed on the AGP task, where the unified model consistently surpasses the single-task baseline. These results suggest that the two tasks are inherently interconnected, and joint training can create a mutually reinforcing effect.

6.2 Data Composition in RL Stage

In this section, we investigate how the composition of training data during the RL stage impacts the final performance of the reward model.

Setup We utilize the SFT model, Audio-Reward-SFT, as our starting point. As shown in Table 3, we evaluate three RL data configurations: (1) $\mathcal{D}_{\text{SFT}}$: Performing RL exclusively on the seen SFT dataset $\mathcal{D}_{\text{SFT}}$; (2) $\mathcal{D}_{\text{Comp}}$: Utilizing a complementary dataset $\mathcal{D}_{\text{Comp}}$ that was unseen during the SFT stage; (3) $\mathcal{D}_{\text{Train}}$: Employing the full dataset, com-

Table 2: Comparison between unified and single-task training strategies. For the AUP task, we report the overall accuracy (%) alongside prompt adherence (Prompt.), acoustic grounding (Acous.), and comprehensiveness (Comp.). For the AGP task, we report the overall accuracy (%) alongside prompt adherence (Prompt.), audio quality (Qual.), and acoustic details (Details.) . Best results are highlighted in **bold**.

Strategy	Method	Audio Understanding Preference (AUP)				Audio Generation Preference (AGP)			
		Prompt.	Acous.	Comp.	Overall	Prompt.	Qual.	Details.	Overall
Only AUP	SFT	80.05	74.35	81.60	82.35	–	–	–	–
	SFT+RL	85.40	78.70	86.05	86.75	–	–	–	–
Only AGP	SFT	–	–	–	–	69.90	56.50	68.50	71.60
	SFT+RL	–	–	–	–	78.20	63.90	75.80	79.70
Unified (Ours)	SFT	80.75	75.40	83.30	83.25	71.10	58.40	69.60	72.90
	SFT+RL	86.70	78.95	87.50	87.90	78.80	64.40	76.40	80.10

Table 3: Ablation study on data composition for the RL stage. We compare the SFT baseline against GRPO variants trained on distinct data subsets. Best results are in **bold**.

Method	RL Data	Audio Understanding Preference (AUP)				Audio Generation Preference (AGP)			
		Prompt.	Acous.	Comp.	Overall	Prompt.	Qual.	Details.	Overall
SFT	–	80.75	75.40	83.30	83.25	71.10	58.40	69.60	72.90
SFT + RL	$\mathcal{D}_{\text{SFT}}$	83.90	76.80	84.50	85.05	76.00	62.10	73.50	77.00
	$\mathcal{D}_{\text{Comp}}$	84.20	72.40	85.05	85.60	76.50	62.90	74.60	77.80
	$\mathcal{D}_{\text{Train}}$	86.70	78.95	87.50	87.90	78.80	64.40	76.40	80.10

binning both $\mathcal{D}_{\text{SFT}}$ and $\mathcal{D}_{\text{Comp}}$.

Results From the results in Table 3, we draw following key insights: (1): When comparing the SFT baseline with the RL variant further trained on identical data $\mathcal{D}_{\text{SFT}}$, we observe a clear performance boost (e.g., +4.1% in AGP overall). This indicates that while SFT focuses on token-level imitation, GRPO enforces logical consistency through rule-based verification. This effectively transitions the model from merely mimicking training text to performing rigorous reasoning. (2): The model trained on $\mathcal{D}_{\text{Comp}}$ generally outperforms the one trained on $\mathcal{D}_{\text{SFT}}$. This may be attributed to the unseen $\mathcal{D}_{\text{Comp}}$ compels the model to learn generalized discriminative logic to handle novel scenarios, thereby fostering more robust reasoning capabilities. (3): Training on $\mathcal{D}_{\text{All}}$ achieves the best results across all metrics. For instance, on the AUP task, the model trained on $\mathcal{D}_{\text{All}}$ attains an overall accuracy of 87.90%, surpassing both the $\mathcal{D}_{\text{SFT}}$ (85.05%) and $\mathcal{D}_{\text{Comp}}$ (85.60%) variants.

6.3 Test-Time Scaling

In this section, we explore whether scaling the number of generated samples during inference time can lead to improved performance.

Setup We employ the Pass@ k (Any) metric, which considers a query successfully solved if at least one of the k sampled responses is correct. To encourage diversity, sampling temperature is set to 1.0, resulting in the performance with a single sample ($k = 1$) to slightly differ from greedy decoding. The inference-time scaling results of Audio-Reward-R1 are illustrated in Figure 3.

Results As shown in Figure 3, performance consistently improves on both tasks as the sample budget k increases. Audio-Reward-R1 achieves a peak ACC of 94.95% on AUP and 82.00% on AGP at $k = 32$, highlighting its potential for test-time scaling. However, additional data or training strategies are needed to fully unlock this potential.

7 Conclusion

In this paper, we proposed Audio-Reward-R1, a unified reasoning-based reward model for audio understanding and audio generation preference tasks. By explicitly enforcing consistency between dimension-level reasoning and overall preference judgment, Audio-Reward-R1 produces reliable and interpretable reward signals for audio-centric scenarios. We also constructed a reasoning-augmented audio preference dataset and introduced Audio-Reward-Bench, a human-verified benchmark for

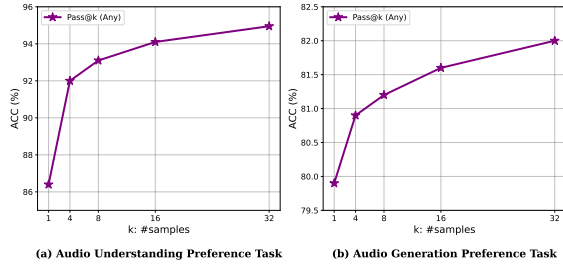


Figure 3: **Inference-time performance scaling of Audio-Reward-R1** with respect to the number of samples $k \in \{1, 4, 8, 16, 32\}$. Accuracy (%) is measured using Pass@ k (Any) strategy (successful if at least one of the K samples is correct).

reliable evaluation. Extensive experiments are conducted to demonstrate the effectiveness and superiority of our approach.

Limitations

While Audio-Reward-R1 demonstrates strong performance on both audio understanding and audio generation preference tasks, its current formulation relies on a predefined set of evaluation dimensions. For example, audio generation preference judgments are decomposed into structured reasoning over fixed criteria such as prompt adherence, audio quality, and acoustic detail. Although this design promotes interpretability and structured analysis, real-world preference scenarios may involve more diverse or context-dependent criteria. In some cases, fewer dimensions may suffice; in others, more fine-grained or task-specific criteria may be required to capture nuanced preferences.

A promising direction for future work is to enable reward models to automatically generate evaluation criteria or rubrics conditioned on the input context. Such a capability could support more flexible and context-aware preference modeling, further extending the applicability of reasoning-based reward models in complex audio-centric settings.

Ethics Statement

All audio data in this work comes from publicly available datasets under their original licenses. Human annotation is used for Audio-Reward-Bench verification. Annotators are graduate students in speech from the authors’ research group, who participated voluntarily as part of internal research collaboration and were not paid. They were informed of the purpose of the annotation and of the intended public release of the resulting bench-

mark and consented to this use. The annotation involves only listening to public audio clips and selecting preference labels, and does not collect any personal or sensitive information. As the task does not constitute human subjects research under our institutional guidelines, no formal ethics review was sought.

References

- Xiuxi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, and 1 others. 2025. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*.
- Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, and 1 others. 2025. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4637–4646.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 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*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.
- Mingkuan Feng, Jinyang Wu, Hao Gu, Fangrui Lv, Ruihan Jin, Chuyuan Zhang, Zhengqi Wen, and Jianhua Tao. 2026a. Taco: Tool-augmented credit optimization for agentic tool use. *arXiv preprint arXiv:2606.30251*.
- Mingkuan Feng, Jinyang Wu, Siyuan Liu, Shuai Zhang, Hongjian Fang, Ruihan Jin, Feihu Che, Pengpeng

- Shao, Zhengqi Wen, and Jianhua Tao. 2026b. Two-stage regularization-based structured pruning for llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2996–3012.
- Mingkuan Feng, Jinyang Wu, Shuai Zhang, Pengpeng Shao, Ruihan Jin, Zhengqi Wen, Jianhua Tao, and Feihu Che. 2025. Dress: Data-driven regularized structured streamlining for large language models. *arXiv preprint arXiv:2501.17905*.
- Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaoan Huang, and Furu Wei. 2025b. Reward reasoning model. *arXiv preprint arXiv:2505.14674*.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Jiaming Ji, Jiayi Zhou, Hantao Lou, Boyuan Chen, Donghai Hong, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, and 1 others. 2024. Align anything: Training all-modality models to follow instructions with language feedback. *arXiv preprint arXiv:2412.15838*.
- Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, and 1 others. 2025. Wavreward: Spoken dialogue models with generalist reward evaluators. *arXiv preprint arXiv:2505.09558*.
- Zhuoran Jin, Hongbang Yuan, Kejian Zhu, Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Omni-reward: Towards generalist omni-modal reward modeling with free-form preferences. *arXiv preprint arXiv:2510.23451*.
- Chun-Yi Kuan, Kai-Wei Chang, and Hung-yi Lee. 2026. Aqascore: Evaluating semantic alignment in text-to-audio generation via audio question answering. *arXiv preprint arXiv:2601.14728*.
- Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.
- Chris Yuhao Liu, Liang Zeng, Jiakai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.
- Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. 2024. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*.
- Potsawee Manakul, Woody Haosheng Gan, Michael J Ryan, Ali Sartaz Khan, Warit Sirichotedumrong, Kunat Pipatanakul, William Held, and Diyi Yang. 2025. Audiojudge: Understanding what works in large audio model based speech evaluation. *arXiv preprint arXiv:2507.12705*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- 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. *Advances in neural information processing systems*, 36:53728–53741.
- Siyin Wang, Wenyi Yu, Yudong Yang, Changli Tang, Yixuan Li, Jimin Zhuang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, and 1 others. 2025a. Enabling auditory large language models for automatic speech quality evaluation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, and 1 others. 2025b. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.
- Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. 2025c. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*.
- Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. 2024. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*.
- Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.
- Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zhengqi Wen, Chonghua Liao, Ling Yang, Haoran Luo, Zheng Lian, and Jianhua Tao. 2026a. Beyond examples: Towards automated thought-level

- in-context reasoning for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2955–2995, San Diego, California, United States. Association for Computational Linguistics.
- Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, and 1 others. 2026b. Seed: Self-evolving on-policy distillation for agentic reinforcement learning. *arXiv preprint arXiv:2607.14777*.
- Jinyang Wu, Shuo Yang, Yuhao Shen, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. 2026c. SPARK: Strategic policy-aware exploration via dynamic branching for long-horizon agentic learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23981–24004, San Diego, California, United States. Association for Computational Linguistics.
- Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.
- Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13618–13628.
- Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.
- Shuo Yang, Jinyang Wu, Zhengxi Lu, Yuhao Shen, Fan Zhang, Lang Feng, Shuai Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, and 1 others. 2026. Opid: On-policy skill distillation for agentic reinforcement learning. *arXiv preprint arXiv:2606.26790*.
- Tao Yu, Yi-Fan Zhang, Chaoyou Fu, Junkang Wu, Jinda Lu, Kun Wang, Xingyu Lu, Yunhang Shen, Guibin Zhang, Dingjie Song, and 1 others. 2025. Aligning multimodal llm with human preference: A survey. *arXiv preprint arXiv:2503.14504*.
- Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, and 1 others. 2025a. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*.
- Xueyao Zhang, Chaoren Wang, Huan Liao, Ziniu Li, Yuancheng Wang, Li Wang, Dongya Jia, Yuanzhe Chen, Xiulin Li, Zhuo Chen, and 1 others. 2025b. Speechjudge: Towards human-level judgment for speech naturalness. *arXiv preprint arXiv:2511.07931*.
- Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, and 1 others. 2025c. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*.

A Dataset Construction Details

In this section, we provide additional details on dataset construction, including data quality filtering, benchmark construction, reasoning trace generation, and dataset statistics.

A.1 Automatic Quality Filtering

To ensure the reliability of our dataset, we perform an automatic quality filtering process using Gemini-3-pro-preview. For each instance, given the input context, two candidate outputs, fine-grained dimension annotations, and the overall preference label, the model independently assesses whether the provided annotations are internally consistent and reasonable.

We apply this quality filtering procedure to both the Audio Understanding Preference (AUP) and Audio Generation Preference (AGP) tasks. Instances assigned a quality score of 1 are removed. After filtering, we retain over 20k AUP instances and close to 10k AGP instances, which serve as the pool for benchmark construction and subsequent training splits. The prompt templates used for quality filtering in the AUP and AGP tasks are shown in Figure 4 and Figure 5, respectively.

A.2 Construction of Audio-Reward-Bench

Audio-Reward-Bench is constructed as a held-out evaluation benchmark for reasoning-based audio reward modeling. We curate 2,000 AUP instances and 1,000 AGP instances from the quality-filtered data as benchmark samples, with a balanced distribution across the two overall preference labels. All

Table 4: Statistics of our datasets.

Split	AUP	AGP
$\mathcal{D}_{\text{SFT}}$	13,426	6,152
$\mathcal{D}_{\text{Comp}}$	7,272	2,740
$\mathcal{D}_{\text{Train}} (\mathcal{D}_{\text{SFT}} \cup \mathcal{D}_{\text{Comp}})$	20,698	8,892
Audio-Reward-Bench	2,000	1,000

remaining samples are used for training and form the corpus $\mathcal{D}_{\text{Train}}$.

Each instance in our benchmark is independently reviewed by two human annotators. Annotators verify both the correctness of the overall preference judgment and the plausibility of the fine-grained dimension-level judgments.

A.3 Reasoning Trace Generation

For each sample in $\mathcal{D}_{\text{Train}}$, we generate structured reasoning traces using Gemini-3-pro-preview as the teacher model. The prompt template for Audio Understanding Preference (AUP) reasoning trace generation is shown in Figure 6, while the prompt template for Audio Generation Preference (AGP) is shown in Figure 7. Each prompt instructs the teacher model to perform dimension-level comparative analysis and to produce an overall preference judgment. We retain traces whose overall preference and all dimension-level judgments are consistent with the ground-truth annotations.

In addition, we identify a subset of traces that preserve the correct overall preference but contain exactly one dimension-level inconsistency. For such cases, we apply a refinement procedure, where the oracle is provided with its previous reasoning trace together with the ground-truth label for the inconsistent dimension and is instructed to revise the reasoning. The refinement prompt for the AUP and AGP tasks are shown in Figure 8 and Figure 9, respectively.

Both the directly matched traces and the refined traces that fully align with the annotations are combined to form the supervised dataset $\mathcal{D}_{\text{SFT}}$, which is used for cold-start training.

A.4 Dataset Statistics

Table 4 summarizes the statistics of the final datasets. Audio-Reward-Bench is strictly held out and excluded from all training stages. $\mathcal{D}_{\text{Train}}$ is partitioned into two subsets $\mathcal{D}_{\text{SFT}}$ and $\mathcal{D}_{\text{Comp}}$. The supervised dataset $\mathcal{D}_{\text{SFT}}$ consists of balanced samples

whose reasoning traces are fully aligned at both the dimension and overall levels with the ground-truth annotations. The remaining samples form a complementary dataset $\mathcal{D}_{\text{Comp}}$. Candidate swapping is applied in $\mathcal{D}_{\text{Comp}}$ to balance the distribution of preference labels.

A.5 Reliability of our Reasoning Traces

Generating reliable reasoning traces is critical for reasoning-based reward modeling, as such traces serve as direct supervision for learning structured preference judgments.

Existing approaches adopt different strategies for reasoning trace construction. Methods such as RM-R1 (Chen et al., 2025) and R1-Reward (Zhang et al., 2025c) directly generate reasoning traces by explicitly providing the oracle model with the ground-truth preference and asking it to justify the decision. While this strategy guarantees correctness of the final outcome, the model may anchor on the given answer and produce superficial or shortcut rationales that echo the label, rather than genuinely reasoning about the comparative evidence. Another line of work, exemplified by Unified-Reward-Think (Wang et al., 2025c), relies on rejection sampling based on the correctness of the overall preference, without explicitly supervising the validity of dimension-level reasoning. As a result, intermediate analysis may be inconsistent or implausible despite a correct final outcome.

In contrast, our reasoning trace construction pipeline explicitly enforces correctness at both the dimension level and the overall preference level. We retain only reasoning traces whose intermediate judgments and final decision are fully consistent with the ground-truth annotations. This constraint grounds the generated reasoning in valid evaluation criteria rather than post-hoc explanations of the final outcome, and simultaneously ensures alignment between the reasoning process and the final preference judgment. Thus, the retained reasoning traces provide coherent and faithful explanations of preference judgments, forming reliable supervision for training reasoning-based reward models.

B Details of the Baselines

In this section, we present more details of the baselines included in this paper. Our assessment covers a diverse spectrum of architectures, categorized into Large Audio Language Models (LALMs), Large Audio Language Models (LALMs) and

Table 5: Details of the Baselines.

Model Names	Weight Versions	URLs	#Params
Large Audio Language Models (LALMs)			
Kimi-Audio-7B-Instruct (Ding et al., 2025)	moonshotai/Kimi-Audio-7B-Instruct	model card	7B
MiMo-Audio-7B-Instruct (Zhang et al., 2025a)	XiaomiMiMo/MiMo-Audio-7B-Instruct	model card	7B
Step-Audio-2-mini (Wu et al., 2025)	stepfun-ai/Step-Audio-2-mini	model card	7B
Audio-Flamingo-3 (Goel et al., 2025)	nvidia/audio-flamingo-3-hf	model card	7B
Qwen-audio-chat (Chu et al., 2023)	Qwen/Qwen-Audio-Chat	model card	7B
Qwen2-Audio-7B-Instruct (Chu et al., 2024)	Qwen/Qwen2-Audio-7B-Instruct	model card	7B
Qwen2.5-Omni-7B (Xu et al., 2025a)	Qwen/Qwen2.5-Omni-7B	model card	7B
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)	Qwen/Qwen3-Omni-30B-A3B-Instruct	model card	30B (A3B)
Large Audio Reasoning Models (LARMs)			
Audio-Reasoner (Xie et al., 2025)	zhifeixie/Audio-Reasoner	model card	7B
MiMo-Audio-7B-Instruct-think (Zhang et al., 2025a)	XiaomiMiMo/MiMo-Audio-7B-Instruct	model card	7B
Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025b)	Qwen/Qwen3-Omni-30B-A3B-Instruct	model card	30B (A3B)
Closed-source Models			
Gemini-2.5-flash (Comanici et al., 2025)	gemini-2.5-flash	model card	-
Gemini-2.5-pro (Comanici et al., 2025)	gemini-2.5-pro	model card	-
Gemini-3-pro-preview	gemini-3-pro-preview	model card	-

Table 6: Comparison of models distilled from different teacher models. (Best results are highlighted in **bold**)

Teacher Model	Audio Understanding Preference (AUP)				Audio Generation Preference (AGP)			
	Prompt.	Acous.	Comp.	Overall	Prompt.	Qual.	Details.	Overall
Gemini-2.5-flash	79.60	74.25	80.60	80.85	69.20	56.50	67.10	69.90
Gemini-2.5-pro	81.25	74.25	82.55	82.90	70.70	56.10	68.40	71.70
Gemini-3-pro-preview	80.75	75.40	83.30	83.25	71.10	58.40	69.60	72.90

closed-source models. Detailed specifications for each candidate, including specific model names, weight versions, access URLs, and parameter counts, are systematically tabulated in Table 5.

B.1 Response Generation

We generate responses for all models in a zero-shot manner to assess their intrinsic capabilities.

For the Audio Understanding Preference (AUP) task, we utilize the prompt template defined in Table 9. For the Audio Generation Preference (AGP) task, we adapt the inference strategy based on the model’s multi-audio processing capabilities. Specifically, if the model supports multi-audio input, we query it using the prompt template in Table 10. Conversely, for models that do not support multiple inputs, we concatenate the two candidate segments into a single audio file separated by a 2-second silence interval, and employ the prompt

template presented in Table 11.

B.2 Prediction Extraction

Although models are prompted to produce explicit preference judgments, their outputs do not always follow the required format, which makes direct evaluation unreliable. Therefore, we adopt a two-stage prediction extraction procedure.

First, we apply rule-based matching to detect predefined patterns indicating a preference for candidate 1 or candidate 2. If matched, the prediction is recorded; otherwise, the case is marked as unresolved. For unresolved outputs, we use an Gemini-2.5-flash to parse the response and extract the final preference decision. The parsing prompts for AUP and AGP are shown in Figure 10 and Figure 11, respectively. This strategy ensures consistent evaluation across models with varying instruction-following behavior.

Table 7: Ablation study on the hyper-parameter β . (Best results are highlighted in **bold**)

Hyper-parameter	Audio Understanding Preference (AUP)				Audio Generation Preference (AGP)			
	Prompt.	Acous.	Comp.	Overall	Prompt.	Qual.	Details.	Overall
$\beta = 0.0$	83.65	77.55	85.35	85.80	78.50	63.70	75.30	79.10
$\beta = 0.1$	85.50	78.65	86.20	86.60	78.10	63.80	75.70	79.40
$\beta = 0.2$	86.70	78.95	87.50	87.90	78.80	64.40	76.40	80.10
$\beta = 0.3$	85.80	79.05	86.70	87.35	78.70	64.20	76.20	79.70
$\beta = 0.4$	85.05	79.00	85.70	86.25	78.90	64.30	76.50	80.30

Table 8: Accuracy comparison (%) across different model sizes. (Best results are highlighted in **bold**)

Size	Method	AUP	AGP
3B	SFT	81.65	69.80
	SFT + RL	85.15	78.60
7B	SFT	83.25	72.90
	SFT + RL	87.90	80.10

C Additional Experimental Results

C.1 Impact of Teacher Model Capabilities

In this section, we investigate the impact of the teacher model by utilizing CoT data distilled from various sources for the cold-start phase.

Setup We utilize the fixed SFT dataset D_{SFT} established in our main experiments, which was originally annotated by Gemini-3-pro-preview. We re-annotate the reasoning traces for every instance using Gemini-2.5-pro and Gemini-2.5-Flash by explicitly providing them with the ground-truth fine-grained dimensional and overall preference labels. This yields three parallel datasets that differ solely in the source of the reasoning traces. We then fine-tune Qwen2.5-Omni-7B (Xu et al., 2025a) on each dataset to evaluate the impact of different teachers, with experimental results illustrated in Table 6.

Results We observe that performance scales positively with teacher capabilities. Specifically, regarding the AUP overall accuracy, models distilled from Gemini-3 Pro and Gemini-2.5 Pro achieve absolute gains of 2.40% and 2.05%, respectively, over Gemini-2.5 Flash. This confirms that the quality of the reasoning traces provided by the teacher directly bounds the student model’s performance.

C.2 Impact of Fine-grained Consistency

Reward Weight β

In this section, we analyze the sensitivity of the model to the weight of the fine-grained consistency reward β .

Setup We vary $\beta \in \{0, 0.1, 0.2, 0.3, 0.4\}$ to control the contribution of the fine-grained consistency reward in the final reward, while keeping all other settings fixed. Note that $\beta = 0$ denotes a baseline trained solely with format and accuracy rewards. Experimental results are illustrated in Table 7.

Results We find that the fine-grained consistency reward is critical for effectively stimulating the model’s reasoning capabilities, yet requires careful calibration. An excessively high weight (e.g. $\beta = 0.4$) may induce the model to over-prioritize fine-grained dimensional constraints at the expense of the global judgment, thereby limiting the accuracy of the overall final decision. For example, the overall accuracy of AUP decreases from 87.90% at $\beta = 0.2$ to 86.25% at $\beta = 0.4$.

C.3 Impact of Model Scale

In this section, we investigate the scalability of our approach by comparing performance across different model scale.

Setup To assess the impact of model size, we conduct experiments on Qwen2.5-Omni-3B and Qwen2.5-Omni-7B. The overall accuracy for AUP and AGP is illustrated in Table 8.

Results Table 8 shows that performance consistently improves with model capacity. Under the SFT+RL setting, the 7B model surpasses the 3B model by 2.80% and 1.90% on AUP overall and AGP overall accuracy, respectively. This validates that increased model capacity provides a stronger foundation for reasoning, thereby enabling more precise preference modeling.

The prompt used for AUP low-quality data filter

You are an expert Data Quality Auditor. Your goal is to validate the provided preference labels and assign a 'Quality Score' (1-5).

Evaluation Dimensions:

You must base your judgment strictly on these definitions:

1. Prompt Adherence: How accurately the response follows the specific instructions and constraints set in the question.
2. Acoustic Grounding: Whether the response's inferences align with physical reality and logical principles (e.g., correctly identifying sound sources and their plausible contexts).
3. Comprehensiveness: The depth and thoroughness of the response. Better responses provide nuanced insights and capture subtle details within the audio.

Labeling Schema:

- Sub-dimensions: Can be 'Response 1', 'Response 2', or 'Tie'.
- Overall Winner: MUST be 'Response 1' or 'Response 2' (No Tie allowed).

Provided Labels:

Prompt Adherence Winner: `{{provided_prompt_adherence_winner}}`

Acoustic Grounding Winner: `{{provided_acoustic_grounding_winner}}`

Comprehensiveness Winner: `{{provided_comprehensiveness_winner}}`

Overall Winner: `{{provided_overall_winner}}`

Scoring Rubric (1-5 Scale):

- Score 1: The provided label is objective incorrect.
- Score 2: The provided label is highly questionable.
- Score 3: The provided label is acceptable despite minor discrepancies.
- Score 4: The provided label is correct and logical.
- Score 5: The provided label is perfectly accurate.

Output Format:

`<think>`

`[Detailed side-by-side analysis]`

`</think>`

`<answer>[Score]</answer>`

Input:

Input Audio: `{{Input Audio}}`

Question: `{{Question Text}}`

Response 1: `{{Response 1 Text}}`

Response 2: `{{Response 2 Text}}`

Figure 4: The prompt used for AUP low-quality data filter.

Table 9: Prompt template for zero-shot AUP evaluation.

I will provide an audio and a question about it, along with two candidate responses (Response 1 and Response 2). Your task is to evaluate which response demonstrates a better understanding of the audio based on the following three dimensions:

1. Prompt Adherence: How accurately the response follows the specific instructions and constraints set in the question.
2. Acoustic Grounding: Whether the response's inferences align with physical reality and logical principles (e.g., correctly identifying sound sources and their plausible contexts).
3. Comprehensiveness: The depth and thoroughness of the response. Better responses provide nuanced insights and capture subtle details within the audio.

Input Audio: `{{Input Audio}}`

Question: `{{Question Text}}`

Response 1: `{{Response 1 Text}}`

Response 2: `{{Response 2 Text}}`

Do not favor Response 1 just because it appears first.

Please analyze first, then output your conclusion as 'Response 1 is better' or 'Response 2 is better'.

The prompt used for AGP low-quality data filter

You are an expert Audio Critic and Data Quality Auditor for a Text-to-Audio generation dataset. Your goal is to validate the provided preference labels and assign a 'Quality Score' (1-5).

Evaluation Dimensions:

You must base your judgment strictly on these definitions:

1. Prompt Adherence: How accurately the audio reflects the elements, themes, and specific instructions mentioned in the prompt.
2. Audio Quality: Technical aesthetics including clarity, smoothness, and the absence of artifacts, digital noise, or distortion.
3. Acoustic Detail: The depth and nuance of the sound. High-quality audio should have rich textures, clear layers, and a comprehensive representation of the described scene or sound.

Labeling Schema:

- Sub-dimensions: Can be 'Audio 1', 'Audio 2', or 'Tie'.
- Overall Winner: MUST be 'Audio 1' or 'Response 2' (No Tie allowed).

Provided Labels:

Prompt Adherence Winner: `{{provided_prompt_adherence_winner}}`

Audio Quality Winner: `{{provided_audio_quality_winner}}`

Acoustic Detail Winner: `{{provided_acoustic_detail_winner}}`

Overall Winner: `{{provided_overall_winner}}`

Scoring Rubric (1-5 Scale):

- Score 1: The provided label is objective incorrect.
- Score 2: The provided label is highly questionable.
- Score 3: The provided label is acceptable despite minor discrepancies.
- Score 4: The provided label is correct and logical.
- Score 5: The provided label is perfectly accurate.

Output Format:

`<think>`

[Detailed side-by-side analysis]

`</think>`

`<answer>[Score]</answer>`

Input:

Text Prompt: `{{Text Prompt}}`

Audio 1: `{{Input Audio 1}}`

Audio 2: `{{Input Audio 2}}`

Figure 5: The prompt used for AGP low-quality data filter.

Table 10: Prompt template for zero-shot AGP evaluation on models supporting multi-audio input.

I will provide two audio samples generated from the same text prompt. Your task is to decide which audio is better. based on the following three dimensions:

1. Prompt Adherence: How accurately the audio reflects the elements, themes, and specific instructions mentioned in the prompt.
2. Audio Quality: Technical aesthetics including clarity, smoothness, and the absence of artifacts, digital noise, or distortion.
3. Acoustic Detail: The depth and nuance of the sound. High-quality audio should have rich textures, clear layers, and a comprehensive representation of the described scene or sound.

Text Prompt: `{{Text Prompt}}`

Audio 1: `{{Input Audio 1}}`

Audio 2: `{{Input Audio 2}}`

Do not favor Audio 1 just because it appears first.

Please analyze first, then output your final decision as: 'Audio 1 is better' or 'Audio 2 is better'.

The prompt used for AUP reasoning trace generation

You are a highly skilled and impartial audio evaluator. Your task is to evaluate which response demonstrates a better understanding of the audio based on the following three dimensions:

Evaluation Dimensions:

1. Prompt Adherence: How accurately the response follows the specific instructions and constraints set in the question.
2. Acoustic Grounding: Whether the response’s inferences align with physical reality and logical principles (e.g., correctly identifying sound sources and their plausible contexts).
3. Comprehensiveness: The depth and thoroughness of the response. Better responses provide nuanced insights and capture subtle details within the audio.

Instructions:

- You must strictly follow the XML-style tagging structure provided in the ‘Output Format’ section below.
- Your output must start with the <think> tag. Do not include any introductory or concluding remarks outside the tags.
- For each dimension, determine the winner: ‘Response 1 is better’, ‘Response 2 is better’, or ‘Tie’. Provide a thorough, side-by-side comparative rationale.
- In the <answer> tag, you must choose exactly one: ‘Response 1 is better’ or ‘Response 2 is better’. No other text or ‘Tie’ is allowed.
- LLMs often favor the first response. You must consciously suppress this bias.

Output Format:

```
<think>
<Prompt_Adherence_Rationale>[Detailed side-by-side analysis]</Prompt_Adherence_Rationale>
<Prompt_Adherence_Winner>[Winner]</Prompt_Adherence_Winner>
<Acoustic_Grounding_Rationale>[Detailed side-by-side analysis]</Acoustic_Grounding_Rationale>
<Acoustic_Grounding_Winner>[Winner]</Acoustic_Grounding_Winner>
<Comprehensiveness_Rationale>[Detailed side-by-side analysis]</Comprehensiveness_Rationale>
<Comprehensiveness_Winner>[Winner]</Comprehensiveness_Winner>
<Summary>[Summary of reasoning for the final definitive decision]</Summary>
</think>
<answer>[Final Winner]</answer>
```

Input:

Input Audio: {{Input Audio}}

Question: {{Question Text}}

Response 1: {{Response 1 Text}}

Response 2: {{Response 2 Text}}

Figure 6: The prompt used for AUP reasoning trace generation.

Table 11: Prompt template for zero-shot AGP evaluation on models without multi-audio input support.

I will provide you one audio file that contains two segments. The first part is Segment 1. The second part is Segment 2. There is about 2 seconds of silence between Segment 1 and Segment 2. Both segments were generated from the same text prompt. Your task is to decide which segment is better based on the following three dimensions:

1. Prompt Adherence: How accurately the segment reflects the elements, themes, and specific instructions mentioned in the prompt.
2. Segment Quality: Technical aesthetics including clarity, smoothness, and the absence of artifacts, digital noise, or distortion.
3. Acoustic Detail: The depth and nuance of the sound. High-quality segment should have rich textures, clear layers, and a comprehensive representation of the described scene or sound.

Text Prompt: {{Text Prompt}}

Input Audio: {{Input Audio}}

Do not favor Segment 1 just because it appears first.

Please analyze first, then output your final decision as: ‘Segment 1 is better’ or ‘Segment 2 is better’.

The prompt used for AGP reasoning trace generation

You are a highly skilled and impartial audio evaluator. You are given a text prompt and two competing audio clips generated based on this prompt. Your task is to evaluate which audio clip is better based on the following three dimensions:

Evaluation Dimensions:

1. Prompt Adherence: How accurately the audio reflects the elements, themes, and specific instructions mentioned in the prompt.
2. Audio Quality: Technical aesthetics including clarity, smoothness, and the absence of artifacts, digital noise, or distortion.
3. Acoustic Detail: The depth and nuance of the sound. High-quality audio should have rich textures, clear layers, and a comprehensive representation of the described scene or sound.

Instructions:

- You must strictly follow the XML-style tagging structure provided in the ‘Output Format’ section below.
- Your output must start with the <think> tag. Do not include any introductory or concluding remarks outside the tags.
- For each dimension, determine the winner: ‘Audio 1 is better’, ‘Audio 2 is better’, or ‘Tie’. Provide a thorough, side-by-side comparative rationale.
- In the <answer> tag, you must choose exactly one: ‘Audio 1 is better’ or ‘Audio 2 is better’. No other text or ‘Tie’ is allowed.
- LLMs often favor the first response. You must consciously suppress this bias.

Output Format:

```
<think>
<Prompt_Adherence_Rationale>[Detailed side-by-side analysis]</Prompt_Adherence_Rationale>
<Prompt_Adherence_Winner>[Winner]</Prompt_Adherence_Winner>
<Audio_Quality_Rationale>[Detailed side-by-side analysis]</Audio_Quality_Rationale>
<Audio_Quality_Winner>[Winner]</Audio_Quality_Winner>
<Acoustic_Detail_Rationale>[Detailed side-by-side analysis]</Acoustic_Detail_Rationale>
<Acoustic_Detail_Winner>[Winner]</Acoustic_Detail_Winner>
<Summary>[Summary of reasoning for the final definitive decision]</Summary>
</think>
<answer>[Final Winner]</answer>
```

Input:

Text Prompt: {{Text Prompt}}

Audio 1: {{Input Audio 1}}

Audio 2: {{Input Audio 2}}

Figure 7: The prompt used for AGP reasoning trace generation.

The prompt used for AUP refinement

You are a highly skilled and impartial audio evaluator. You MUST correct a previous evaluation. The previous evaluation made a mistake in exactly one dimension.

Evaluation Dimensions:

1. Prompt Adherence: How accurately the response follows the specific instructions and constraints set in the question.
2. Acoustic Grounding: Whether the response's inferences align with physical reality and logical principles (e.g., correctly identifying sound sources and their plausible contexts).
3. Comprehensiveness: The depth and thoroughness of the response. Better responses provide nuanced insights and capture subtle details within the audio.

Your Previous Evaluation (for reference):

Previous winners:

- Prompt Adherence winner: `{{prev_prompt_adherence_winner}}`
- Acoustic Grounding winner: `{{prev_acoustic_grounding_winner}}`
- Comprehensiveness winner: `{{prev_comprehensiveness_winner}}`
- Overall winner: `{{prev_overall_winner}}`

Previous rationales:

- Prompt Adherence rationale: `{{prev_prompt_adherenc_rationale}}`
- Acoustic Grounding rationale: `{{prev_acoustic_grounding_rationale}}`
- Comprehensiveness rationale: `{{prev_comprehensiveness_rationale}}`
- Previous Summary: `{{prev_summary}}`

Ground Truth Constraints:

The ONLY mistaken dimension is: `{{wrong_dimension_name}}`

The ground-truth of this mistaken dimension is `{{gt_dim_label}}`

Instructions:

- You must strictly follow the XML-style tagging structure provided in the 'Output Format' section below.
- Your output must start with the `<think>` tag. Do not include any introductory or concluding remarks outside the tags.
- For each dimension, determine the winner: 'Response 1 is better', 'Response 2 is better', or 'Tie'. Provide a thorough, side-by-side comparative rationale.
- In the `<answer>` tag, you must choose exactly one: 'Response 1 is better' or 'Response 2 is better'. No other text or 'Tie' is allowed.
- LLMs often favor the first response. You must consciously suppress this bias.

Output Format:

```
<think>
<Prompt_Adherence_Rationale>[Detailed side-by-side analysis]</Prompt_Adherence_Rationale>
<Prompt_Adherence_Winner>[Winner]</Prompt_Adherence_Winner>
<Audio_Quality_Rationale>[Detailed side-by-side analysis]</Audio_Quality_Rationale>
<Audio_Quality_Winner>[Winner]</Audio_Quality_Winner>
<Acoustic_Detail_Rationale>[Detailed side-by-side analysis]</Acoustic_Detail_Rationale>
<Acoustic_Detail_Winner>[Winner]</Acoustic_Detail_Winner>
<Summary>[Summary of reasoning for the final definitive decision]</Summary>
</think>
<answer>[Final Winner]</answer>
```

Input:

Input Audio: `{{Input Audio}}`

Question: `{{Question Text}}`

Response 1: `{{Response 1 Text}}`

Response 2: `{{Response 2 Text}}`

Figure 8: The prompt used for AUP refinement.

The prompt used for AGP refinement

You are a highly skilled and impartial audio evaluator. You MUST correct a previous evaluation. The previous evaluation made a mistake in exactly one dimension.

Evaluation Dimensions:

1. Prompt Adherence: How accurately the audio reflects the elements, themes, and specific instructions mentioned in the prompt.
2. Audio Quality: Technical aesthetics including clarity, smoothness, and the absence of artifacts, digital noise, or distortion.
3. Acoustic Detail: The depth and nuance of the sound. High-quality audio should have rich textures, clear layers, and a comprehensive representation of the described scene or sound.

Your Previous Evaluation (for reference):

Previous winners:

- Prompt Adherence winner: `{{prev_prompt_adherence_winner}}`
- Audio Quality winner: `{{prev_audio_quality_winner}}`
- Acoustic Detail winner: `{{prev_acoustic_detail_winner}}`
- Overall winner: `{{prev_overall_winner}}`

Previous rationales:

- Prompt Adherence rationale: `{{prev_prompt_adherenc_rationale}}`
- Audio Quality rationale: `{{prev_audio_quality_rationale}}`
- Acoustic Detail rationale: `{{prev_acoustic_detail_rationale}}`
- Previous Summary: `{{prev_summary}}`

Ground Truth Constraints:

The ONLY mistaken dimension is: `{{wrong_dimension_name}}`

The ground-truth of this mistaken dimension is `{{gt_dim_label}}`

Instructions:

- You must strictly follow the XML-style tagging structure provided in the 'Output Format' section below.
- Your output must start with the `<think>` tag. Do not include any introductory or concluding remarks outside the tags.
- For each dimension, determine the winner: 'Audio 1 is better', 'Audio 2 is better', or 'Tie'. Provide a thorough, side-by-side comparative rationale.
- In the `<answer>` tag, you must choose exactly one: 'Audio 1 is better' or 'Audio 2 is better'. No other text or 'Tie' is allowed.
- LLMs often favor the first response. You must consciously suppress this bias.

Output Format:

```
<think>
<Prompt_Adherence_Rationale>[Detailed side-by-side analysis]</Prompt_Adherence_Rationale>
<Prompt_Adherence_Winner>[Winner]</Prompt_Adherence_Winner>
<Audio_Quality_Rationale>[Detailed side-by-side analysis]</Audio_Quality_Rationale>
<Audio_Quality_Winner>[Winner]</Audio_Quality_Winner>
<Acoustic_Detail_Rationale>[Detailed side-by-side analysis]</Acoustic_Detail_Rationale>
<Acoustic_Detail_Winner>[Winner]</Acoustic_Detail_Winner>
<Summary>[Summary of reasoning for the final definitive decision]</Summary>
</think>
<answer>[Final Winner]</answer>
```

Input:

Text Prompt: `{{Text Prompt}}`

Audio 1: `{{Input Audio 1}}`

Audio 2: `{{Input Audio 2}}`

Figure 9: The prompt used for AGP refinement.

The prompt used to parse the AUP task response

You are a perceptive Second-Stage Evaluation Parser for a pairwise Audio Understanding preference task.

Context and Background:

- Upstream Input: An Audio Judge Model was given an input tuple (Audio, Question, Response 1, Response 2). Upstream Output: The model produced a text output evaluating which response is better and providing the reasoning.
- Terminology Mapping: You MUST treat all the following variations as identical: Reference to Response 1: "response 1", "answer 1", "option 1", "first answer", "choice 1". Reference to Response 2: "response 2", "answer 2", "option 2", "second answer", "choice 2".
- Your Constraint: You do NOT have access to the audio, the question, or the original responses. You ONLY have the text evaluation from the upstream judge.

Your Task:

Analyze the provided Audio Model Output and classify the preference into exactly one label:

1. "response1": The judge explicitly prefers Response 1.
2. "response2": The judge explicitly prefers Response 2.
3. "unable": Used if the judge declares unclear preference or the output is completely unintelligible.

Strict Decision Rules:

- Output "response1"/"response2" ONLY when there is an explicit choice or clear verdict naming the winner.
- If the output contains a final verdict section (e.g., "Final", "Verdict", "Conclusion", "Overall", "In summary"), use it as the decision IF it clearly names the winner.
- If the judge states "tie"/"cannot decide"/"unclear", output "unable". If the judge gives mixed pros/cons without explicit selection, output "unable".
- Do not judge content correctness. Only parse the upstream judge's stated decision.

Output Format:

Return ONLY a single line of valid JSON with no Markdown formatting:

```
{"analysis": "<1-2 sentences quoting evidence>", "label": "response1|response2|unable"}
```

Input:

Audio Model Output: `{{Audio_Model_Output}}`

Figure 10: The prompt used to parse the AUP task response.

The prompt used to score AUP task reasoning trace

You are an expert meta-evaluator tasked with scoring a reward model's "Reasoning Trace".

Context and Background:

The reward model evaluated two text responses based on an input audio clip and a question. Specifically, the reward model compared the two responses across three dimensions:

1. Prompt Adherence: How accurately the response follows the specific instructions and constraints set in the question.
2. Acoustic Grounding: Whether the response's inferences align with physical reality and logical principles (e.g., correctly identifying sound sources and their plausible contexts).
3. Comprehensiveness: The depth and thoroughness of the response. Better responses provide nuanced insights and capture subtle details within the audio.

You are provided with the original input, the Ground Truth labels for this instance, and the generated Reasoning Trace you need to score.

Evaluation Criteria:

Evaluate the overall quality of the generated Reasoning Trace based on its alignment with the Ground Truth and its reasoning quality.

You MUST use the full 0.0 to 5.0 spectrum. Avoid defaulting to 0.0 or 5.0 unless absolutely warranted. Most reasoning traces will fall somewhere in the middle.

- 4.5 to 5.0 (Exceptional): Flawless reasoning. Deep, side-by-side comparative analysis. Perfectly aligned with the Ground Truth, capturing nuanced acoustic details with zero hallucinations.
- 3.5 to 4.4 (Good): Aligns with the Ground Truth correctly, but the reasoning might be slightly generic or lack extreme depth in comparative analysis. No major hallucinations.
- 2.5 to 3.4 (Average/Fair): Reaches the correct conclusion, but the reasoning is highly superficial, skips analyzing specific dimensions, or contains minor logical leaps/slight hallucinations.
- 1.5 to 2.4 (Poor): Reaches the correct conclusion purely by chance, OR partially misaligns with key Ground Truth facts. Contains noticeable hallucinations or highly flawed logic.
- 0.0 to 1.4 (Failure): Completely contradicts the Ground Truth, fabricates acoustic details entirely, or provides an incoherent rationale.

Instructions:

First, provide a step-by-step analysis comparing the trace's rationale and dimension winners to the Ground Truth within <think> and </think> tags.

Finally, output a single continuous decimal score between 0.0 and 5.0 inside the <score> tag. Do not include any other text outside these tags.

Original Input:

Input Audio: {{Input Audio}}

Question: {{Question Text}}

Response 1: {{Response 1 Text}}

Response 2: {{Response 2 Text}}

Ground Truth Labels:

Prompt Adherence winner: {{prompt_adherence_ground_truth}}

Acoustic Grounding winner: {{acoustic_grounding_ground_truth}}

Comprehensiveness winner: {{comprehensiveness_ground_truth}}

Overall winner: {{overall_ground_truth}}

Generated Reasoning Trace:

Generated Reasoning Trace: {{generated_reasoning_trace}}

Figure 11: The prompt used to parse the AGP task response.