REINFORCEMENT LEARNING WITH EVOLVING RUBRICS AS REWARDS FOR AUDIO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio reasoning is essential for machine understanding of the acoustic world. Reinforcement learning with verifiable rewards can elicit such reasoning, yet existing reward designs are complementary in their limitations: outcome-based rewards supervise only the final answer and let the model reach it without attending to the audio, whereas process-based rewards score the reasoning itself but rely on coarse, hand-crafted, and fixed criteria that neither adapt to each question nor stay grounded in the acoustic evidence. Moreover, questions differ in what they demand, with some hinging on perception and others on multi-step reasoning, and any static criterion weakens as the policy improves. Supervising the reasoning process with fine-grained, audio-grounded, and adaptive rewards is therefore crucial, yet challenging since such rewards are impractical to design by hand for every sample. To this end, we introduce AUDIORUBRICS, a reinforcement learning framework that supervises audio reasoning with evolving, audio-grounded rubric rewards. AUDIORUBRICS synthesizes per-sample rubrics from the raw waveform and, conditioned on the model's own rollouts, regenerates and reweights criteria per group, supplying a continuous learning signal that keeps targeting the current policy's weaknesses as static criteria saturate. Comprehensive evaluations across three audio reasoning benchmarks reveal that AUDIORUBRICS substantially outperforms a wide range of open-source and training-based baselines. Furthermore, our analysis shows that gains depend on the rubric generator and judge's capability, and AUDIORUBRICS converges to a stable reasoning length that avoids both degenerate collapse and unbounded growth. The improvement in audio perception further demonstrates the effectiveness of anchoring supervision in the acoustic evidence.

1 INTRODUCTION

Audio understanding and reasoning is the ability to perceive acoustic signals and infer meaning from them, spanning the spoken word, environmental sound, and music (Ma et al., 2025a; Gong et al., 2024). As a fundamental facet of human intelligence, this ability lets us extract rich information from what we hear—identifying who is speaking and how, recognizing events from sound alone, and grasping the structure and affect of music—and act on it in the world. Similarly, by equipping machines with the capacity to not merely transcribe audio but reason over its acoustic content, we move beyond shallow pattern recognition toward genuine auditory comprehension, thereby enabling systems that support voice interaction (Chen et al., 2025) and a broad range of real-world applications from healthcare (Shah et al., 2026) to multimedia analysis (Polyak et al., 2024).

Recently, reinforcement learning with verifiable rewards (RLVR) has emerged as a powerful tool for eliciting reasoning in large audio language models (LALMs). Outcome-based methods (Li et al., 2025; Rouditchenko et al., 2025) reward only the final answer. While simple and verifiable, they leave the reasoning trace unsupervised, allowing the model to arrive at the correct option without a genuinely sound reasoning process for understanding the audio. In contrast, process-based methods (Zhifei et al., 2025; Wu et al., 2026; Fan et al., 2025) score the reasoning itself, providing a denser learning signal, but they rely on coarse and fixed criteria that assess only the surface quality of the reasoning trace. Critically, such criteria do not verify whether the reasoning is grounded in the audio, resulting in a response can be rewarded for fluent or well-structured text that is not actually supported by the acoustic evidence. Beyond this, applying a single fixed rubric to all questions is fundamentally mismatched to the heterogeneity of audio reasoning, since different questions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of AUDIORUBRICS with existing post-training designs for audio reasoning. Unlike prior approaches that supervise only the final answer or rely on a coarse, hand-crafted, fixed reward signal, AUDIORUBRICS jointly delivers a fine-grained reward that co-evolves with the policy.

Method	Process Supervision	Fine-grained Criteria	Audio-Grounded	Automatic Generation	Evolve with Policy
R1-AQA (Li et al., 2025)	✗	✗	✗	✗	✗
Omni-R1 (Rouditchenko et al., 2025)	✗	✗	✗	✗	✗
Audio-Reasoner (Zhifei et al., 2025)	✓	✗	✓	✗	✗
Audio-Thinker (Wu et al., 2026)	✓	✗	✗	✓	✗
CESAR (Fan et al., 2025)	✓	✓	✗	✗	✗
AUDIORUBRICS	✓	✓	✓	✓	✓

demand different evaluation criteria. Some can be answered from accurate perception alone; for these, soliciting additional reasoning incurs unnecessary computational overhead and increases the risk of hallucination. Others genuinely require multi-step inference, and for these the rubric should place greater weight on the logical soundness of the reasoning chain. A uniform standard rewards both cases identically and therefore fails to match the evaluation to what each question actually requires. Furthermore, a static rubric becomes progressively less informative as the policy improves. Once the model reliably satisfies a fixed criterion, that criterion no longer discriminates between stronger and weaker responses and its training signal saturates, leaving little gradient to drive further improvement. The evaluation standard thus stops scaling with the model's growing capability. These limitations call for an evaluation that is multi-dimensional, question-specific rather than uniform, and capability-adaptive, continually raising the bar as the model improves. However, as summarized in Table 1, existing approaches cover at most a subset, leaving audio reasoning without an evaluation signal that is audio-grounded, question-adaptive, and capability-scaling.

To address these limitations, we introduce AUDIORUBRICS, a reinforcement learning framework that supervises audio reasoning with evolving rubric-based rewards. AUDIORUBRICS first constructs initial rubrics for each question in the training data to capture the basic evaluation criteria, generating them directly from the raw waveform so that every criterion is grounded in the acoustic evidence actually present in the clip. Combined with the verifiable RLVR reward, these rubric-based rewards provide dense process-level supervision. Beyond merely checking the final answer, they assess whether the reasoning trace is logically sound and faithfully grounded in the audio, supplying a learning signal on the reasoning process itself. However, during RL training, the policy steadily improves and satisfies more and more of the initial rubrics. The criteria then become non-discriminative, and the training signal saturates, leaving little gradient to drive further improvement. To resolve this, we condition the rubric generator on the group of rollouts sampled by the policy and compare them against one another. By contrasting where stronger and weaker rollouts diverge, we first elicit new criteria that build on the existing ones to promote more advanced reasoning. We then prune any rubric whose verdicts are identical across all rollouts, since a criterion that every response passes or fails cannot contribute advantage. The rubric set continually retires saturated criteria and adopts harder ones, so the model receives a sustained learning signal whose standard co-evolves with its growing capability. However, the rubric-based reward incentivizes the model to satisfy more criteria by producing ever longer reasoning chains, which can be exploited to hack the reward through redundant, circular, or hallucinated reasoning. To mitigate this, we incorporate an overthinking penalty that linearly penalizes reasoning length, keeping the trace informative but concise.

We evaluate AUDIORUBRICS on three widely used audio reasoning benchmarks: MMAU (Sakshi et al., 2024), MMAR (Ma et al., 2025b), and MMSU (Wang et al., 2025). Experimental results show that AUDIORUBRICS consistently outperforms a wide range of open-source models and RL baselines of comparable model size. Further analysis reveals that the rubrics grow progressively more advanced as the policy improves over training, confirming that the evaluation standard co-evolves with the model, and ablation studies verify the contribution of each component.

2 PRELIMINARY

2.1 GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Given a multimodal input (Q, A) consisting of an audio input A and a textual query Q, the policy model samples G candidate responses $o = \{o_1, \ldots, o_G\}$, and a reward function $r(\cdot)$ assigns a

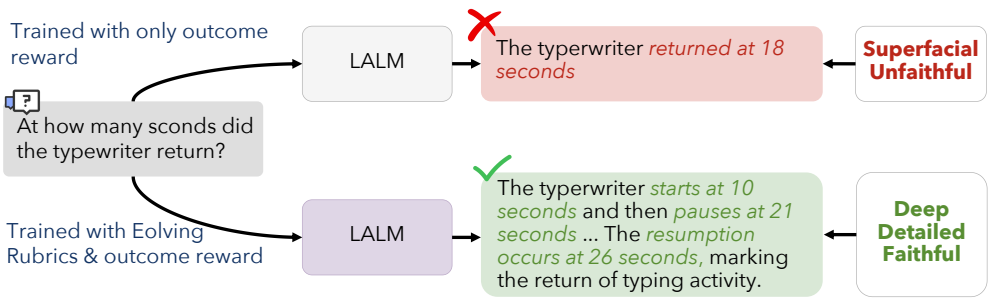


Figure 1: Motivation of the proposed AUDIORUBRICS. The goal is to progressively enhance the reasoning ability of LALMs by rewarding not only the correct outcome but also the reasoning process with evolving rubrics, which continually distill new criteria from the model's own rollouts to push the reasoning from superficial guesses toward deep, faithful, and evidence-grounded explanations.

reward score to each response, yielding $\{r(o_1), \ldots, r(o_G)\}$. GRPO encourages the policy to generate responses that maximize a weighted sum reward $R(o)$, defined by:

$$R(o) = \sum_{i=1}^{G} \frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}}(o_i)} \cdot \underbrace{\frac{r(o_i) - \text{mean}(\{r(o_j)\}_{j=1}^{G})}{\text{std}(\{r(o_j)\}_{j=1}^{G})}}_{\text{Advantage } A_i} \tag{1}$$

where $\pi_\theta(o)$ denotes the probability of LLM generating the response o, and $\pi_{\theta_{\text{old}}}$ represents the policy parameters from a recently optimized state. The latter term is the advantage A_i of the i-th candidate. To ensure training stability and avoid large deviations from the original language model behavior, the final training objective incorporates a KL-divergence regularization term (Guo et al., 2025), penalizing divergence between π_θ and π_{ref}:

$$\max_{\pi_\theta} \mathbb{E}_{o \sim \pi_{\theta_{\text{old}}}(\cdot|Q,A)}[R(o) - \beta_{\text{KL}} \, \text{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})] \tag{2}$$

where β_{KL} is a scaling coefficient. We omit the clipping operation for simplicity.

2.2 RUBRIC AS REWARDS

Rubrics define explicit evaluation criteria for assessing the quality of model responses (Viswanathan et al., 2026; Gunjal et al., 2025). We consider sample-wise rubrics, in which the evaluation criteria are specified on a per-example basis in natural language: given a query x with associated rubrics $\mathcal{R}_x = \{(r_{x,k}, w_{x,k})\}_{k=1}^{K}$, where $r_{x,k}$ denotes a rubric item and $w_{x,k}$ its weight satisfying $\sum_{k=1}^{K} w_{x,k} = 1$, we evaluate a final response y using the rubric-based score

$$S(x, o) = \sum_{k=1}^{K} w_{x,k} \, \text{JUDGE}(x, r_{x,k}, o). \tag{3}$$

Each rubric is evaluated by a judge LM that is conditioned on the query x and outputs $\{0, 1\}$ based on whether the response o satisfies $r_{x,k}$. During training, we optimize the expected rubric score over the training questions using RL. Using rubrics as rewards offers several advantages: their concrete, well-defined items reduce susceptibility to judge model bias and promote objective evaluation, yielding consistent and comparable scores across different LLM-as-a-judge runs.

3 AUDIORUBRICS: RL WITH EVOLVING RUBRICS AS REWARDS

Although RLVR has proven effective in enhancing logical reasoning, it supervises only the final outcome and leaves the reasoning process unchecked. A correct answer certifies neither that the underlying reasoning is grounded in the audio, nor that the reasoning itself is informative (Figure 1). Moreover, such a fixed, outcome-level criterion gradually saturates as training progresses, and the

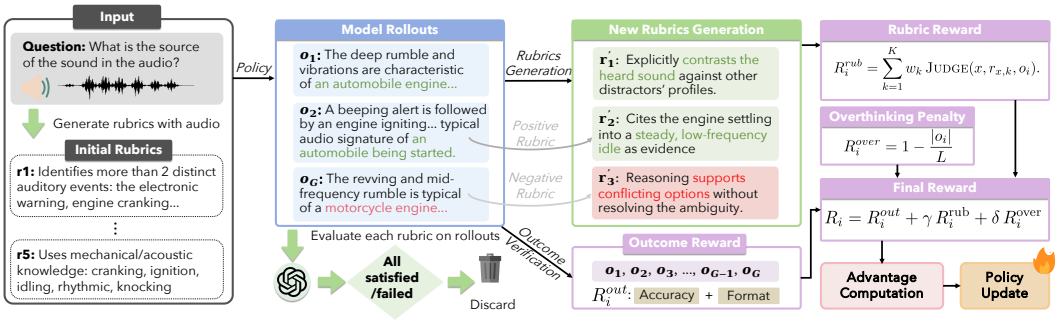


Figure 2: Overview of AUDIORUBRICS. A rubric generator initializes audio-grounded rubrics for each question. At each step, the policy LALM samples rollouts, and the generator elicits refined rubrics from the rollouts and judges every rubric on them; non-discriminative rubrics are discarded. The resulting rubric score is combined with the outcome reward for GRPO optimization.

learning signal weakens accordingly. AUDIORUBRICS addresses both issues by augmenting the outcome reward with an evolving, audio-grounded rubric reward that directly scores the reasoning process. This design not only encourages high-quality reasoning, but also—because the rubrics evolve alongside the policy—keeps the reward signal informative throughout training. As shown in Figure 2, we first initialize audio-grounded rubrics for each question to assess basic response quality. At each RLVR step, a judge model evaluates the rollouts against the current rubric set, elicits refined rubrics from the rollouts themselves, prunes non-discriminative ones, and re-weights the survivors; the resulting rubric score is then combined with the outcome reward for GRPO optimization.

3.1 TASK FORMULATION

In this work, we investigate the audio reasoning task in the context of large audio language models (LALMs). Let $\mathcal{D} = (x_1, x_2, \ldots, x_N)$ be an audio reasoning dataset, where each data sample $x_i = (A, Q, y^\star)$ comprises an audio input A (e.g., speech, sound, or music), a textual query Q, and the corresponding ground-truth answer $y^\star$. The audio reasoning task is defined as follows: given a data sample $x_i \in \mathcal{D}$ as input, the LALM is required to reason over the audio and predict an answer y that matches the ground-truth answer $y^\star$.

3.2 REWARD DESIGN

Outcome Rewards. Let y_i be the answer extracted from the <answer> span of rollout o_i. The accuracy reward is $R_i^{\text{acc}} = \mathbb{I}[\, y_i = y^\star \,]$, and the format reward $R_i^{\text{fmt}} = \mathbb{I}[\, o_i$ contains the required <think>...</think><answer>...</answer> structure$\,]$ enforces the reasoning-then-answer template. Both terms are deterministic, and we denote their combination by

$$R_i^{\text{out}} = \alpha\, R_i^{\text{acc}} + \beta\, R_i^{\text{fmt}}, \tag{4}$$

where R_i^{fmt} enforces structured output while R_i^{acc} drives the reasoning capability. We refer to RLVR trained with $R_i = R_i^{\text{out}}$ as accuracy-only RLVR.

Audio-Grounded Rubric Initialization. Given an audio clip A, a question Q, and its ground-truth answer $y^\star$, we first synthesize an initial set of weighted rubrics that will act as a process-level reward. Instead of passing a textual transcript, we feed the raw waveform of A to a single audio-capable model Φ, which serves as both the rubric generator and the judge throughout AUDIORUBRICS, as native audio input; this keeps every generated criterion anchored to acoustic evidence actually present in the clip. To spread the rubrics across complementary, non-redundant aspects of response quality, we fix a taxonomy of K evaluation dimensions $\mathcal{F} = \{f_1, \ldots, f_K\}$ in advance (e.g., auditory-evidence grounding. See full dimensions in Figure 9) and request exactly one criterion per dimension,

$$\mathcal{R}_0 = \{(r_k, w_k)\}_{k=1}^{K} = \Phi(A, Q, y^\star; \mathcal{F}), \tag{5}$$

where each r_k is a binary, positively phrased merit statement and the importance weights satisfy $w_k \in (0, 1)$ with $\sum_{k=1}^{K} w_k = 1$. The weights are assigned by Φ in the same call that generates the rubrics. $\mathcal{R}_0$ is generated once per question before training.

Evolving Rubrics as Rewards. Though the initial rubrics provide fine-grained criteria, the policy gradually satisfies them as training progresses. Therefore, to obtain a continuous reward signal that targets the weaknesses of the current policy, we design evolving rubrics as rewards. Specifically, at each iteration, conditioned on A, Q, the current rollouts $\{o_i\}_{i=1}^{G}$, and the rubric set $\mathcal{R}_{\text{prev}}$ retained from the previous iteration in which this question was sampled, Φ performs three operations. For the first iteration of a question, $\mathcal{R}_{\text{prev}}$ is initialized to the set $\mathcal{R}_0$. Carrying over the survivors from the previous iteration lets the rubric set accumulate the harder criteria distilled in earlier iterations, so the evaluation standard ratchets upward as the policy improves.

(i) Elicitation and judging. Φ proposes up to N_{new} new rubrics $\mathcal{R}_{\text{new}}$ that are non-redundant with $\mathcal{R}_{\text{prev}}$ and target reasoning quality and audio grounding. Because these rubrics are induced from observed rollouts, they also carry negative criteria (satisfied is a flaw), capturing recurring failure modes that positively phrased initial rubrics cannot express. Φ then returns binary judgments $j_{k,i} \in \{0, 1\}$ of whether rollout o_i satisfies rubric r_k for every $r_k \in \mathcal{R} = \mathcal{R}_{\text{prev}} \cup \mathcal{R}_{\text{new}}$ and every rollout. Polarity normalization flips negative-polarity judgments so that $b_{k,i} = 1$ always denotes a good outcome:

$$b_{k,i} = \begin{cases} j_{k,i}, & r_k \text{ is positive rubric,} \\ 1 - j_{k,i}, & r_k \text{ is negative rubric,} \end{cases} \tag{6}$$

yielding comparable binary vectors $\{b_{k,i}\}_{i=1}^{G}$ across rubrics.

(ii) Variance filtering. For each rubric, we compute the within-group standard deviation $s_k = \text{std}_i\{b_{k,i}\}_{i=1}^{G}$ and prune every rubric with $s_k = 0$. Such a rubric is either satisfied by all rollouts, indicating that the policy has already mastered this criterion, or by none, indicating that it lies beyond the policy's current capability. In both cases, it assigns every rollout the same verdict, and thus contributes nothing to the group-relative advantage. Among the survivors, we rank rubrics by s_k and retain the top M most discriminative ones or all, if fewer than M survive. We denote the kept set of rubrics by $\mathcal{K}$.

(iii) Weighting and scoring. Conditioned on A, Q, and $\mathcal{K}$, Φ re-assigns positive weights $\{w_k\}_{k\in\mathcal{K}}$ with $\sum_{k\in\mathcal{K}} w_k = 1$. The weights are re-assigned by Φ for this question, and the final rubric reward is:

$$R_i^{\text{rub}} = \sum_{k\in\mathcal{K}} w_k \, b_{k,i} \in [0, 1]. \tag{7}$$

This elicitation, filtering, and re-weighting is performed at every training step, independently for each prompt group. Because rubrics are thus updated per group conditioned on the actual rollouts, R_i^{rub} stays informative and provides a more advanced reward signal for the current policy.

Overthinking Penalty. Although the rubric reward improves reasoning quality, satisfying more criteria favors longer reasoning, which risks degenerating into redundant or circular traces that accumulate errors and hallucinations (Mahmoud et al., 2026). To alleviate this, we regularize reasoning length with a penalty:

$$R_i^{\text{over}} = 1 - \frac{|o_i|}{L}, \tag{8}$$

where $|o_i|$ is the token length of the reasoning trace o_i and L is a reference length that normalizes the penalty and sets its scale relative to the other reward terms. Taken in isolation this term favors shorter traces, but in combination it counterbalances the rubric reward, which pushes toward longer traces that satisfy more criteria. The rubric term thus improves reasoning quality with richer details, while the length term bounds its quantity, and their equilibrium keeps the trace informative but concise.

Overall Reward and Optimization. The three reward terms above play complementary roles, in which the outcome reward anchors final-answer correctness, the rubric reward densely supervises the reasoning process, and the overthinking penalty bounds its length. The per-rollout reward is their weighted combination:

$$R_i = R_i^{\text{out}} + \gamma \, R_i^{\text{rub}} + \delta \, R_i^{\text{over}}, \tag{9}$$

where γ controls the impact of the rubric reward and δ weights the overthinking penalty. Accuracy-only RLVR is recovered at $\gamma = \delta = 0$. During training, we replace the reward function $r(\cdot)$ in Eq. 1 with our evolving rubric reward R_i and train the LALM to maximize the GRPO objective in Eq. 2. Algorithm 1 in the appendix summarizes the per-group computation of the evolving rubric reward.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Benchmarks. We evaluate on three audio reasoning benchmarks. MMAU Test-mini (Sakshi et al., 2024) has 1000 multiple-choice questions over 27 tasks spanning speech, sound, and music. MMAR (Ma et al., 2025b) provides 1000 QA pairs from real-world videos that mix the three modalities, organized into four reasoning layers for deep reasoning. MMSU (Wang et al., 2025) adds 5000 triplets over 47 spoken-language tasks probing fine-grained paralinguistic and phonological cues such as prosody, emotion, and speaker traits. Together they cover sound, music, and speech and their mixtures, from perception to multi-step reasoning.

Baselines and Evaluation Metrics. We compare AUDIORUBRICS against a comprehensive suite of baselines spanning proprietary models, open-source LALMs, and recent training-based audio reasoning methods, including 4 proprietary models and 13 open-source models. For proprietary models, we evaluate GPT-4o-Audio (OpenAI, 2024), GPT-audio-1.5 (OpenAI, 2026), and Gemini-3-Flash, Gemini-3.1-Pro (Google DeepMind, 2025). Open-source LALMs include Qwen2-Audio (7B) (Chu et al., 2024), Phi-4-Multimodal (5.6B) (Abdin et al., 2024), Kimi-Audio (7B) (Ding et al., 2025), Step-Audio-2-mini (7B) (Wu et al., 2025), Qwen2.5-Omni (7B) (Xu et al., 2025), MiMo-Audio (7B) (Zhang et al., 2025), and Audio-Flamingo 3 (7B) (Ghosh et al., 2026). For training-based reasoning methods, we further compare with Audio-Reasoner (7B) (Zhifei et al., 2025), R1-AQA (7B) (Li et al., 2025), Omni-R1 (7B) (Rouditchenko et al., 2025), Ke-Omni-R (7B) (Zhao et al., 2025), Audio-Thinker (7B) (Wu et al., 2026), and CESAR (7B) (Fan et al., 2025), all of which apply RL or SFT-based post-training to elicit audio reasoning. For evaluation, we set temperature $\tau = 0$ to perform greedy decoding for all models and use accuracy as the main evaluation metric.

Training details. The training data is drawn from the AVQA dataset (Yang et al., 2022), which is also used by most training-based baselines (R1-AQA, Omni-R1, Ke-Omni-R, Audio-Thinker, and CESAR). Following R1-AQA (Li et al., 2025), we extract audio from videos and construct audio-text pairs by replacing "video" with "audio" in the questions, resulting in 40,176 training samples. All training runs use 4 H100 GPUs. We use Gemini-3.1-Pro as the rubric judge and generator. Our experiments employ Qwen2.5-Omni-7B, sampling $G = 8$ responses per training example. See more implementation details in Appendix A.

4.2 MAIN RESULTS

We present the performance comparison between AUDIORUBRICS and existing powerful models and methods across three widely used audio understanding and reasoning benchmarks in Tables 2 and 3. Based on the results, we have the following key observations:

(i) AUDIORUBRICS achieves superior performance on all benchmarks among all models and methods of similar size. AUDIORUBRICS attains the best overall accuracy on all three benchmarks among similar-size models. The advantage is also evident at the dimension level. AUDIORUBRICS ranks first in MMAU Sound across all models, and leads all similar-size models in the Speech category of both benchmarks. AUDIORUBRICS also leads or matches the best similar-size baseline on three of the four mixed-modality splits of MMAR. These consistent gains at both the overall and the per-dimension level demonstrate the effectiveness of our approach, which provides a richer learning signal for training.

(ii) The audio perception capability of AUDIORUBRICS shows tangible improvements. Specifically, on the perception split of MMSU, AUDIORUBRICS surpasses the best-performing baseline of comparable size by 4.3 points, an 8.9% relative improvement, with consistent gains across all three perception dimensions. This indicates that our approach teaches the model to anchor its reasoning in the acoustic evidence it actually hears, thereby translating process-level supervision into a genuine gain in perception and ensuring that correct answers rest on faithfully perceived evidence.

(iii) AudioRubrics narrows the gap to proprietary models. AudioRubrics improves the Qwen2.5-Omni-7B by 12.8, 9.1, and 5.3 points on MMAU, MMAR, and MMSU, respectively. On MMAU, AUDIORUBRICS even surpasses GPT-4o-Audio, GPT-audio-1.5, Gemini-3-Flash, and Gemini-3.1-Pro. On MMAR, it outperforms GPT-4o-Audio. This suggests that training with rubric-based

Table 2: Results on the MMSU and MMAU Test-mini benchmarks. The top two results are highlighted in **bold** and underline among non-proprietary models, respectively. Audio-Thinker is not evaluated on MMSU as its checkpoint is not publicly released.

Models	MMSU (%↑)									MMAU Test-mini (%↑)			
	Perception				Reasoning				Avg	Sound	Music	Speech	Avg
	Seman.	Phono.	Para.	Avg	Seman.	Phono.	Para.	Avg	All				
Proprietary Models													
GPT-4o-Audio	59.70	41.56	21.44	39.67	80.83	78.74	26.25	71.96	56.38	64.56	56.29	66.67	62.50
GPT-audio-1.5	82.20	64.06	29.11	54.84	90.61	65.71	32.54	72.52	63.40	78.08	63.17	83.48	74.90
Gemini-3-Flash	80.94	68.98	65.45	70.54	92.06	90.38	57.01	86.53	78.28	74.47	76.65	81.38	77.50
Gemini-3.1-Pro	90.08	76.79	79.41	81.09	88.36	87.41	61.19	84.21	82.60	75.98	75.45	81.38	77.60
Open-source Models													
Qwen2-Audio	52.14	32.87	35.56	39.02	77.62	64.81	46.67	68.90	53.27	67.27	56.29	55.26	59.60
Phi-4-Multimodal	38.72	34.86	29.56	33.41	57.81	65.94	42.09	57.59	44.96	65.47	64.37	67.27	65.70
Kimi-Audio	57.64	42.30	35.74	43.52	81.77	76.65	55.22	76.03	59.28	75.68	66.77	62.16	68.20
Step-Audio-2-mini	38.90	31.55	41.19	37.13	70.04	71.44	44.18	67.02	51.60	79.30	68.44	68.16	72.73
Qwen2.5-Omni-7B	55.12	37.33	39.35	42.50	88.00	**81.37**	48.36	79.83	60.57	69.07	59.58	66.97	65.20
MiMo-Audio-7B	57.80	43.74	38.22	45.04	73.74	66.73	38.21	65.99	55.18	68.47	**82.58**	73.65	74.90
Audio-Flamingo 3	–	–	–	–	–	–	–	–	62.30	79.88	76.55	66.37	74.26
Training-based models													
Audio-Reasoner	46.77	35.19	34.16	37.64	75.00	64.38	37.01	65.45	51.10	67.87	69.16	66.07	67.70
R1-AQA	55.12	35.19	37.72	41.09	79.15	66.22	51.04	70.04	55.10	68.77	64.37	63.66	65.60
Omni-R1	58.43	44.49	42.48	47.13	88.18	77.79	46.87	78.26	62.20	81.38	68.86	73.57	74.60
Ke-Omni-R	58.74	46.31	40.50	47.09	86.82	74.31	**60.00**	78.06	62.08	79.28	70.06	74.47	74.60
Audio-Thinker	–	–	–	–	–	–	–	–	–	77.48	70.36	73.37	73.70
CESAR	60.16	50.16	39.50	48.45	88.72	80.66	57.01	**81.07**	64.24	83.48	73.05	74.77	77.10
AUDIORUBRICS	**66.14**	**51.34**	**44.16**	**52.75**	**89.17**	81.26	49.25	80.45	**65.86**	**85.89**	72.16	**75.98**	**78.00**

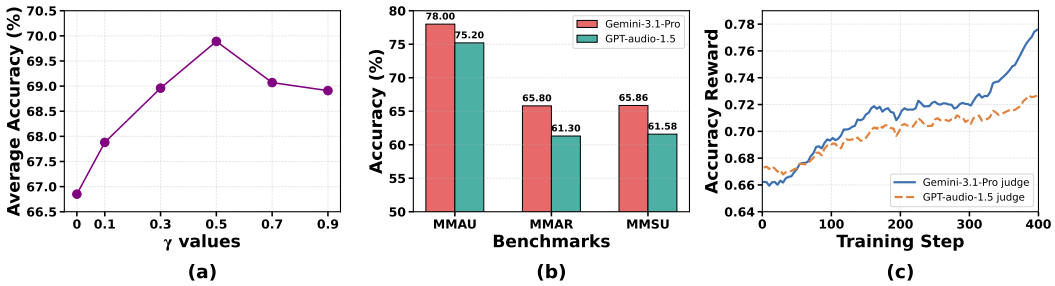


Figure 3: (a) Average performance across all benchmarks with varying rubric reward weight γ. (b) Comparison of accuracy across benchmarks when using different judging and reward generators. (c) Dynamics of accuracy reward during training when using different judging and reward generators.

evolving supervision thus allows to approach, and in some cases exceed, much larger proprietary models.

4.3 Further Analysis on Reward Design and Judgments

Effect of the rubric reward weight. Table 4 and Figure 3 (a) vary the rubric weight γ with the overthinking penalty fixed. Across the entire range, adding the rubric reward consistently improves over GRPO. As γ increases, AUDIORUBRICS receives a stronger process-level reward and its accuracy improves accordingly, peaking at $\gamma = 0.5$. Beyond this point, however, performance declines since an overly large rubric weight downweights the accuracy reward, so the policy is increasingly optimized toward satisfying the rubric criteria rather than producing the cor-

Table 4: **Impact of rubric reward weighting.** Performance on the audio benchmarks with different rubric weights. $\Delta^{\%}_{rel}$ is relative to the GRPO baseline.

Rubric	MMAU		MMAR		MMSU	
	ACC	$\Delta^{\%}_{rel}$	ACC	$\Delta^{\%}_{rel}$	ACC	$\Delta^{\%}_{rel}$
GRPO	75.20	–	62.20	–	63.14	–
$\gamma = 0.10$	77.10	↑ 2.53	62.80	↑ 0.96	63.74	↑ 0.95
$\gamma = 0.30$	77.30	↑ 2.79	64.90	↑ 4.34	64.68	↑ 2.44
$\gamma = 0.50$	**78.00**	↑ 3.72	**65.80**	↑ 5.79	65.86	↑ 4.31
$\gamma = 0.70$	77.70	↑ 3.32	63.40	↑ 1.93	**66.12**	↑ 4.72
$\gamma = 0.90$	77.80	↑ 3.46	64.10	↑ 3.05	64.84	↑ 2.69

Table 3: Results on the MMAR benchmark. Top two results are highlighted in **bold** and underline.

Method	Single Modality			Mixed Modality				Overall
	Sound	Music	Speech	Mix-S-M	Mix-S-Sp	Mix-M-Sp	Mix-S-M-Sp	
Proprietary Models								
GPT-4o-Audio	53.94	50.97	70.41	63.64	72.48	62.20	75.00	63.50
GPT-audio-1.5	63.64	53.40	81.29	63.64	83.49	73.17	70.83	72.00
Gemini-3-Flash	66.06	62.62	85.71	90.91	83.03	75.61	75.00	76.10
Gemini-3.1-Pro	70.91	65.53	84.01	90.91	84.86	79.27	66.67	77.50
Large Audio Language Models (LALMs)								
Qwen2-Audio	33.33	24.27	32.31	9.09	31.19	30.49	25.00	30.00
Phi-4-Multimodal	32.73	31.07	50.68	18.18	41.74	57.32	41.67	41.70
Kimi-Audio	–	–	–	–	–	–	–	48.00
Step-Audio-2-mini	51.52	32.04	49.66	45.45	51.83	54.88	62.50	47.50
Qwen2.5-Omni-7B	58.79	40.78	59.86	54.55	61.93	67.07	58.33	56.70
MiMo-Audio-7B	52.73	37.38	62.59	36.36	69.27	59.76	62.50	56.70
Audio-Flamingo 3	–	–	–	–	–	–	–	60.10
Training-based models								
Audio-Reasoner	43.64	33.50	32.99	45.45	42.66	31.71	25.00	36.80
R1-AQA	52.73	40.78	49.32	9.09	50.92	52.44	50.00	48.30
Omni-R1	59.39	50.49	61.56	54.55	60.09	63.41	41.67	58.20
Ke-Omni-R	63.64	47.09	62.93	63.64	68.35	67.07	45.83	60.90
Audio-Thinker	67.27	53.88	64.29	**72.73**	71.56	**73.17**	66.67	65.30
CESAR	66.06	**55.83**	62.24	63.64	67.43	60.98	66.67	62.70
AUDIORUBRICS	**68.48**	51.94	**68.71**	**72.73**	**72.02**	65.85	**70.83**	**65.80**

rect answer, ultimately weakening the very signal that grounds the reasoning in a verifiably correct outcome.

Effect of the overthinking penalty. Table 5 studies the length-penalty weight δ. When the weight δ is small, the policy tends to generate overlong reasoning traces, since it seeks to satisfy more rubrics, which weakens the impact of the accuracy reward. As δ increases, the performance peaks at $\delta = 0.15$, which effectively controls overthinking. However, a larger penalty would excessively compress the reasoning trace, weakening reasoning ability and leading to suboptimal performance. See Appendix C.1 for the length dynamics during training.

Table 5: **Impact of overthinking penalty weighting.** Performance on the audio benchmarks with different overthinking penalty weights. $\Delta^{\%}_{rel}$ is relative to the GRPO baseline.

OT	MMAU		MMAR		MMSU	
	ACC	$\Delta^{\%}_{rel}$	ACC	$\Delta^{\%}_{rel}$	ACC	$\Delta^{\%}_{rel}$
GRPO	75.20	–	62.20	–	63.14	–
$\delta = 0.05$	77.40	↑ 2.93	63.70	↑ 2.41	64.30	↑ 1.84
$\delta = 0.10$	77.70	↑ 3.32	**66.10**	↑ 6.27	65.56	↑ 3.83
$\delta = 0.15$	**78.00**	↑ 3.72	65.80	↑ 5.79	**65.86**	↑ 4.31
$\delta = 0.20$	77.20	↑ 2.66	64.40	↑ 3.54	64.70	↑ 2.47

Effect of the Rubric Generator and Judge Model. To examine how the choice of generator and judge affects AUDIORUBRICS, we replace Gemini-3.1-Pro with the weaker GPT-audio-1.5 as both. As shown in Figure 3 (b), performance drops with GPT-audio-1.5, even falling below the vanilla GRPO baseline. The training dynamics in Figure 3 (c) corroborate this: after a brief warm-up in which the two runs are comparable, the Gemini-3.1-Pro policy pulls ahead and stays superior, whereas the GPT-audio-1.5 variant learns more slowly and plateaus lower. We attribute this gap to the lower-quality rubrics from the weaker model, whose imprecise or poorly grounded criteria yield noisy rewards, indicating that AUDIORUBRICS hinges on a sufficiently capable generator and judge.

4.4 ABLATION STUDY

We conduct ablation studies to analyze the contribution of each component in AUDIORUBRICS. As shown in Table 6, compared to the base model, RL training substantially improves audio reasoning capability. Building on RL training, the introduction of rubric rewards yields a further gain, where the initial rubrics provide fine-grained criteria that jointly assess the quality of the reasoning process and the accuracy of audio perception. Allowing the rubrics to evolve, rather than relying solely on

Table 6: Ablation results of different components in AUDIORUBRICS.

Method	Components				Overall Accuracy (%)		
	RL Training	Static Rubrics	Evolving Rubrics	Overthinking Penalty	MMAU	MMAR	MMSU
Full Method (AUDIORUBRICS)	✓	✓	✓	✓	78.00	65.80	65.86
Ablating Overthinking Penalty	✓	✓	✓	✗	77.20	64.60	65.22
Ablating Evolving Rubrics	✓	✓	✗	✗	76.20	63.60	65.44
Ablating Static Rubrics	✓	✗	✗	✗	75.20	62.20	63.14
Ablating RL Training	✗	✗	✗	✗	65.20	56.70	60.57

their initialized form, further improves performance, as evolving rubrics continually target the current model's reasoning weaknesses and supply progressively more advanced criteria that drive further gains in reasoning ability. The overthinking penalty mitigates the tendency to generate excessively long reasoning traces merely to satisfy the rubrics at the expense of answer accuracy.

4.5 RUBRIC EVOLUTION ANALYSIS

To understand how the rubric set adapts during training, we track the ratio of newly evolved rubrics that are adopted for the reward at each step. As shown in Figure 4, this share grows steadily from roughly 35% early in training to about 60% by the end. In the early stage, the policy mainly satisfies the initial rubrics, so few new criteria are needed. As the model improves and satisfies the easier rubrics, the variance filter increasingly prunes them, and a growing fraction of the supervision comes from newly generated rubrics that provide more advanced, fine-grained criteria. These results demonstrate that the rubric set continually upgrades to track the model's current weaknesses, rather than supervising against a frozen rubric list. See Appendix B.2 for a qualitative analysis.

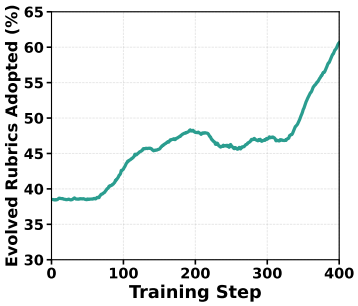


Figure 4: The ratio of newly evolved rubrics adopted during training.

5 RELATED WORK

Audio Understanding and Reasoning. Early work learns audio representations by adapting CLIP to audio (Radford et al., 2021; Wu et al., 2022). More recent Large Audio Language Models (LALMs) integrate audio into large language model backbones and, after large-scale pretraining (Ghosh et al., 2025; Chu et al., 2023; Xu et al., 2025; Chu et al., 2024; Ghosh et al., 2024), sharpen reasoning through supervised fine-tuning on reasoning traces (Zhifei et al., 2025; Li et al., 2026), tool use (Chen et al., 2026a; Tong et al., 2026; Lee et al., 2025), or reinforcement learning (Fan et al., 2025; Tian et al., 2025; Wu et al., 2026; Yu et al., 2026a). Unlike these methods, which reward only the final answer or score reasoning against fixed, hand-crafted criteria, AUDIORUBRICS supervises audio reasoning with evolving, audio-grounded rubrics tailored to each question.

Rubric as Rewards. Reinforcement learning with verifiable rewards excels where correctness is programmatically checkable (Guo et al., 2025; Yu et al., 2024; 2026b;c), but its scalar reward does not extend to open-ended tasks that demand multi-criteria judgment. Rubric-based rewards fill this gap by scoring responses against fine-grained, instance-specific criteria (Bi et al., 2025; Raghavendra et al., 2026; Mu et al., 2024; Zhou et al., 2025; Tyagi et al., 2026; Jacob Dineen et al., 2025), which are elicited from a strong model and synthesized offline or online from policy rollouts (Kim et al., 2024; Gupta et al., 2025; Liu et al., 2025; Xie et al., 2025; Shen et al., 2026; Shao et al., 2025; Sheng et al., 2026; Rezaei et al., 2025; Xu et al., 2026), and have recently been extended to vision-language and omni-modal reasoning (Jia et al., 2025; Yu et al., 2026d; Chen et al., 2026b; Qiu et al., 2026; Kong et al., 2026). Unlike these text- and vision-centric efforts, we introduce evolving rubrics for audio reasoning, providing a fine-grained and adaptive learning signal to train LALMs to reason.

6 CONCLUSION

We introduce AudioRubrics, an RL framework that supervises audio reasoning with audio-grounded rubrics evolved from the model's own rollouts to keep rewards informative, and that substantially outperforms a wide range of baselines on audio understanding and reasoning benchmarks.

AI USE DISCLOSURE

We used generative AI tools solely to polish the writing of this paper. The authors reviewed and edited all AI-assisted text and take full responsibility for the content.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Baolong Bi, Shenghua Liu, Yiwei Wang, Siqian Tong, Lingrui Mei, Yuyao Ge, Yilong Xu, Jiafeng Guo, and Xueqi Cheng. Reward and guidance through rubrics: Promoting exploration to improve multi-domain reasoning. *arXiv preprint arXiv:2511.12344*, 2025.

Liyang Chen, Hongkai Chen, Yujun Cai, Sifan Li, Qingwen Ye, and Yiwei Wang. Audiorouter: Data efficient audio understanding via rl based dual reasoning. *arXiv preprint arXiv:2602.10439*, 2026a.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, et al. Slam-omni: Timbre-controllable voice interaction system with single-stage training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2262–2282, 2025.

Yukun Chen, Jiaming Li, Longze Chen, Ze Gong, Jingpeng Li, Zhen Qin, Hengyu Chang, Ancheng Xu, Zhihao Yang, Hamid Alinejad-Rokny, et al. Rucl: Stratified rubric-based curriculum learning for multimodal large language model reasoning. *arXiv preprint arXiv:2602.21628*, 2026b.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *CoRR*, abs/2311.07919, 2023. doi: 10.48550/ARXIV.2311.07919. URL https://doi.org/10.48550/arXiv.2311.07919.

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.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

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.

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

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S. Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *CoRR*, abs/2503.03983, 2025. doi: 10.48550/ARXIV.2503.03983. URL https://doi.org/10.48550/arXiv.2503.03983.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886, 2026.

Yuan Gong, Hongyin Luo, Alexander Liu, Leonid Karlinsky, and James R Glass. Listen, think, and understand. In *International Conference on Learning Representations*, volume 2024, pp. 18516–18545, 2024.

Google DeepMind. Gemini 3.1 pro. https://deepmind.google/models/gemini/pro/, 2025. Preview model. Knowledge cutoff January 2025. Accessed: 2026-06-28.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

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

Taneesh Gupta, Shivam Shandilya, Xuchao Zhang, Rahul Madhavan, Supriyo Ghosh, Chetan Bansal, Huaxiu Yao, and Saravan Rajmohan. Carmo: Dynamic criteria generation for context aware reward modelling. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2202–2261, 2025.

Aswin RRV Jacob Dineen, Qin Liu, Zhikun Xu, Xiao Ye, Ming Shen, Zhaonan Li, Shijie Lu, Chitta Baral, Muhao Chen, and Ben Zhou. Qa-lign: Aligning llms through constitutionally decomposed qa. *arXiv preprint arXiv:2506.08123*, 2025.

Mengzhao Jia, Zhihan Zhang, Ignacio Cases, Zheyuan Liu, Meng Jiang, and Peng Qi. Autorubric-r1v: Rubric-based generative rewards for faithful multimodal reasoning. *arXiv preprint arXiv:2510.14738*, 2025.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024.

Zicheng Kong, Dehua Ma, Zhenbo Xu, Alven Yang, Yiwei Ru, Haoran Wang, Zixuan Zhou, Fuqing Bie, Liuyu Xiang, Huijia Wu, et al. Omni-rrm: Advancing omni reward modeling via automatic rubric-grounded preference synthesis. *arXiv preprint arXiv:2602.00846*, 2026.

Kuan-Yi Lee, Tsung-En Lin, and Hung-Yi Lee. Audio-maestro: Enhancing large audio-language models with tool-augmented reasoning. *arXiv preprint arXiv:2510.11454*, 2025.

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.

Longhao Li, Hongjie Chen, Zehan Li, Qihan Hu, Jian Kang, Jie Li, Lei Xie, and Yongxiang Li. Audio-cogito: Towards deep audio reasoning in large audio language models. *arXiv preprint arXiv:2604.12527*, 2026.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. *arXiv preprint arXiv:2510.07743*, 2025.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *CoRR*, abs/2501.07246, 2025a. doi: 10.48550/ARXIV.2501.07246. URL https://doi.org/10.48550/arXiv.2501.07246.

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

Anas Mahmoud, MohammadHossein Rezaei, Zihao Wang, Anisha Gunjal, Bing Liu, and Yunzhong He. Reward hacking in rubric-based reinforcement learning. *arXiv preprint arXiv:2605.12474*, 2026.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. *Advances in Neural Information Processing Systems*, 37:108877–108901, 2024.

OpenAI. GPT-4o Audio (gpt-4o-audio-preview). https://developers.openai.com/api/docs/models/gpt-4o-audio-preview, 2024.

OpenAI. gpt-audio-1.5. https://developers.openai.com/api/docs/models/gpt-audio-1.5, 2026.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Weijie Qiu, Dai Guan, Junxin Wang, Zhihang Li, Yongbo Gai, Mengyu Zhou, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Rationale matters: Learning transferable rubrics via proxy-guided critique for vlm reward models. *arXiv preprint arXiv:2603.16600*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Mohit Raghavendra, Anisha Gunjal, Bing Liu, and Yunzhong He. Agentic rubrics as contextual verifiers for swe agents. *arXiv preprint arXiv:2601.04171*, 2026.

MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Bing Liu, Yunzhong He, and Afra Feyza Akyürek. Online rubrics elicitation from pairwise comparisons. *arXiv preprint arXiv:2510.07284*, 2025.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-r1: Do you really need audio to fine-tune your audio llm? *arXiv preprint arXiv:2505.09439*, 2025.

Sakshi 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.

Meet Shah, Jason Gusdorf, Anil Palepu, Chunjong Park, Jack W O'Sullivan, Vishnu Ravi, Tim Strother, Pavel Dubov, Aliya Rysbek, Toshiyuki Fukuzawa, et al. Towards conversational medical ai with eyes, ears and a voice. *arXiv preprint arXiv:2605.09272*, 2026.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

William F Shen, Xinchi Qiu, Chenxi Whitehouse, Lisa Alazraki, Shashwat Goel, Francesco Barbieri, Timon Willi, Akhil Mathur, and Ilias Leontiadis. Rethinking rubric generation for improving llm judge and reward modeling for open-ended tasks. *arXiv preprint arXiv:2602.05125*, 2026.

Leheng Sheng, Wenchang Ma, Ruixin Hong, Xiang Wang, An Zhang, and Tat-Seng Chua. Reinforcing chain-of-thought reasoning with self-evolving rubrics. *arXiv preprint arXiv:2602.10885*, 2026.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, et al. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

Siqian Tong, Xuan Li, Yiwei Wang, Baolong Bi, Yujun Cai, Shenghua Liu, Yuchen He, and Chengpeng Hao. Autagent: A reinforcement learning framework for tool-augmented audio reasoning. *arXiv preprint arXiv:2602.13685*, 2026.

Utkarsh Tyagi, Xingang Guo, MohammadHossein Rezaei, Daniel George, Anas Mahmoud, Jackson Lee, Bing Liu, and Yunzhong He. Not every rubric teaches equally: Policy-aware rubric rewards for rlvr. *arXiv preprint arXiv:2605.20164*, 2026.

Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Sherry Wu. Checklists are better than reward models for aligning language models. *Advances in Neural Information Processing Systems*, 38:114728–114754, 2026.

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.

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.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4563–4567. IEEE, 2022.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-thinker: Guiding large audio language model when and how to think via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33962–33970, 2026.

Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, et al. Auto-rubric: Learning to extract generalizable criteria for reward modeling. *arXiv e-prints*, pp. arXiv–2510, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215, 2025. doi: 10.48550/ARXIV.2503.20215. URL https://doi.org/10.48550/arXiv.2503.20215.

Ran Xu, Tianci Liu, Zihan Dong, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tao Zhao, and Haoyu Wang. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable llm post-training. *arXiv preprint arXiv:2602.01511*, 2026.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3480–3491, 2022.

Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. Flow of reasoning: Training llms for divergent reasoning with minimal examples. *arXiv preprint arXiv:2406.05673*, 2024.

Fangxu Yu, Zinan Lin, Xiaodong Liu, Weijia Xu, Michael Xu, Tianyi Zhou, and Jianfeng Gao. Weak-to-strong on-policy distillation. *arXiv preprint arXiv:2607.26246*, 2026a.

Fangxu Yu, Ziyao Lu, Liqiang Niu, Fandong Meng, and Jie Zhou. Arrowgev: Grounding events in video via learning the arrow of time. *arXiv preprint arXiv:2601.06559*, 2026b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026c.

Ya-Qi Yu, Fangyu Hong, Xiangyang Qu, Hao Wang, Gaojie Wu, Qiaoyu Luo, Nuo Xu, Huixin Wang, Wuheng Xu, Yongxin Liao, et al. Visual preference optimization with rubric rewards. *arXiv preprint arXiv:2604.13029*, 2026d.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

Shuaijiang Zhao, Tingwei Guo, Cheng Wen, Bajian Xiang, and Wei Zou. Ke-omni-r: Achieving advanced audio reasoning with a concise 50-words think process. https://github.com/shuaijiang/Ke-Omni-R, 2025.

Xie Zhifei, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23840–23862, 2025.

Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, et al. Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning. *arXiv preprint arXiv:2508.16949*, 2025.

Table 7: Training hyperparameters of AUDIORUBRICS.

Hyperparameter	**Value**
Algorithm	GRPO
Base Model	Qwen2.5-Omni-7B
Max Prompt Length	4096
Max Response Length	1024
KL Coefficient (β_{KL})	0.001
Learning Rate	1e-6
LR Scheduler	Constant
Sampling Temperature	1.0
Batch Size	8
Rollout Number (G)	8
Training Steps	400
Precision	bf16
Accuracy / Format Reward Weight	0.9 / 0.1
Rubric Reward Weight (γ)	0.5
Overthinking Penalty Weight (δ)	0.15
Reference Length (L)	256
Rubrics Kept per Group (top-M)	5
New Rubrics per Group	3
Rubric Generator / Judge	Gemini-3.1-Pro

A IMPLEMENTATION DETAILS

We leverage Qwen2.5-Omni-7B (Xu et al., 2025) as our base model. During RLVR training, we optimize the policy with GRPO for 400 steps, and set a batch size of 8, a learning rate 1e-6 with a constant schedule, a number of rollouts $G = 8$ with a sampling temperature 1.0, and KL coefficient $\beta_{\text{KL}} = 0.001$, with a maximum prompt length of 4096 and a maximum response length of 1024 tokens under bf16 precision. For the reward, we set the accuracy and format weights to $\alpha = 0.9$ and $\beta = 0.1$, the rubric reward weight to $\gamma = 0.5$ and the overthinking penalty weight to $\delta = 0.15$, following the sensitivity analyses in Section 4.3, and the reference length to $L = 256$ tokens. Gemini-3.1-Pro serves as the rubric generator and judge Φ, keeping the top $M = 5$ most discriminative rubrics per group and eliciting up to $N_{\text{new}} = 3$ new rubrics at each step. The checkpoint from the final step is used for all evaluations. All experiments were conducted on 4 H100 GPUs. See the summary in Table 7.

B QUALITATIVE ANALYSIS

B.1 CASE STUDY

To understand how our reward design improves audio reasoning, we conduct a qualitative analysis in Figure 5 where the model must infer a speaker's communicative intent. The GRPO baseline answers incorrectly, anchoring on the most lexically salient token in the audio and producing a one-line justification that restates this surface topic without reasoning about the underlying intent. In contrast, AUDIORUBRICS reconstructs the interaction step by step, tracks the evolving dialogue state, and rules out each distractor against the conversational flow rather than against isolated keywords. We attribute this behavior to our evolving rubric-based training process reward, which supplies fine-grained credit to intermediate reasoning steps and thereby incentivizes the model to ground its answer in the full context rather than in a single salient cue.

B.2 COMPARISON BETWEEN INITIAL AND EVOLVED RUBRICS

Figure 6 compares the two rubric sources on a location example. The kept initial rubrics cover easy sub-tasks—grounding the sound and ruling out the water options—that both right and wrong rollouts already pass. The evolved rubrics instead come from the rollouts and target the hard split the answer

Case Study: GRPO and AUDIORUBRICS

Question: Why does the uncle say the last sentence in the audio?

Choices:

A. Because he wants to explain that he is imitating Superman
B. Because he thinks his nephew is praising him
C. Because he feels his nephew doesn't understand his performance
D. Because he thinks Superman is cool

Ground Truth: B

GRPO (Baseline) — Predicted A ✗
The uncle says the last sentence because he wants to explain that he is imitating Superman.

AudioRubrics (Ours) — Predicted B ✓
On careful listening, the audio features a nephew asking to hear an impression, then the uncle begins it (groaning "No, not the kryptonite!"). The nephew corrects him, noting it is Superman, to which the uncle responds "Thank you, I've been practicing a lot." This indicates the uncle says the last sentence because he *mistakenly believes his nephew is praising* his performance. Choices A and D do not match the conversational context, and C is not supported by the dialogue flow.

Figure 5: Qualitative case study. The GRPO baseline latches onto the surface topic (Superman) and answers A, whereas our model tracks the speaker's *intent* across the dialogue and correctly infers that the uncle misreads the correction as praise.

turns on, *field* vs. *woods*, which the generator weights as highly as the grounding rubric. Thus the evolved rubrics add the question-specific discrimination that the fixed set lacks.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 ANALYSIS ON RESPONSE LENGTH

Figure 7 tracks the average response length over the course of training under three reward configurations, revealing two opposite pathologies that our full design avoids. Under vanilla GRPO (orange), the response length steadily collapses: after a brief initial plateau, it decays toward near-zero by the end of training, indicating that the policy learns to shortcut to terse, under-reasoned answers that maximize the outcome reward without producing an intermediate reasoning trace. Removing the overthinking penalty from our objective (blue) induces the opposite failure mode: the response length grows in an unbounded, increasingly unstable manner and ultimately explodes, as the model is rewarded for ever-longer generations regardless of their quality, inflating verbosity and compute cost without commensurate gains in correctness. In contrast, our full method (green) maintains a stable response length throughout training, fluctuating within a narrow band rather than drifting toward either extreme. This stability indicates that the rubric-based process reward and the length regularization act in complementary ways: the former supplies dense credit for substantive reasoning steps, preventing the degenerate collapse seen under GRPO, while the latter discourages uninformative padding, keeping the generated rationales concise yet sufficiently detailed. The result is a training process that converges to a well-calibrated reasoning length, which we find correlates with the accuracy improvements reported in Table 2.

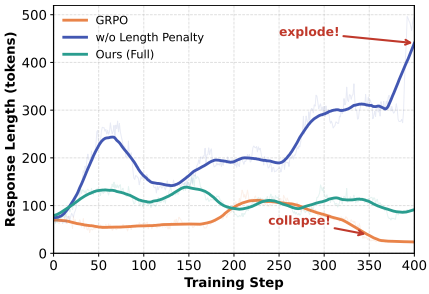


Figure 7: Average response length during training under three reward configurations. Vanilla GRPO collapses toward near-zero length, removing the overthinking penalty leads to unbounded growth, and our full method maintains a stable length throughout.

Case Study: Rubric Co-Evolution

Question: Where does the audio take place?

Choices:

A. aquatic
B. at sea
C. field
D. woods

Ground Truth: C

(Footsteps crunch through dry vegetation in an open outdoor soundscape with faint birdsong—no water sounds.)

Retained Static Rubrics

Two of the five initial criteria still discriminate and are kept, covering the *easy* sub-tasks—grounding the evidence and ruling out the water options:

- **[Auditory Evidence Grounding]** (w=0.35) identifies the footsteps crunching through dry vegetation as the primary evidence.
- **[Distractor Elimination]** (w=0.15) rules out "aquatic"/"at sea" by the complete absence of water-related sounds.

The other three generic static criteria—reasoning clarity, focus, and terminology—have saturated and are pruned by the variance filter.

Newly Evolved Rubrics

The generator distills two criteria that target the discrimination the fixed rubrics never asked for, and gives the harder one the *top* weight:

- **[Woodland vs. Field Disambiguation]** (w=0.35) explicitly differentiates "field" from "woods" by citing specific acoustic differences, such as spatial acoustics (open vs. dense) or the nature of the rustling.
- **[Background Environmental Cues]** (w=0.15) identifies secondary sounds—bird chirping, wind noise—to further corroborate the outdoor terrestrial setting.

Figure 6: Qualitative case study of rubric co-evolution. Two static rubrics are retained for the easy sub-tasks, while the generator adds the fine-grained "field vs. woods" disambiguation—the crux of the question—and weights it as highly as the retained grounding criterion.

Table 8: Results on the 3B model. $\Delta^{\%}_{rel}$ denotes the relative improvement over the GRPO baseline.

Method	MMAU		MMAR		MMSU	
	ACC	$\Delta^{\%}_{rel}$	ACC	$\Delta^{\%}_{rel}$	ACC	$\Delta^{\%}_{rel}$
3B						
GRPO	73.00	–	58.30	–	60.26	–
AUDIORUBRICS	**73.90**	↑1.23	**59.30**	↑1.72	**62.46**	↑3.65

C.2 Robustness to Different Model Scale

We further apply our method to the 3B model to demonstrate its generalizability and robustness across model scales. We present the results in Table 8, comparing against the GRPO baseline trained under the identical setting. From the table, we observe that our method consistently outperforms GRPO across all three benchmarks, improving MMAU, MMAR, and MMSU by 1.23%, 1.72%, and 3.65% relative, respectively. The gains persist even at this smaller scale, where the reduced model capacity leaves less room for reasoning-oriented rewards to take effect. The consistent improvement indicates that our rubric-based process reward provides a robust learning signal across different model scales.

D Prompts

This section lists the full prompts used throughout our pipeline: the inference prompt for benchmark evaluation, the static rubric generation prompt used for per-question initialization, the evolving-

Algorithm 1 Evolving rubric reward for one prompt group

Require: audio A, question Q, ground-truth answer $y^\star$, rollouts $\{o_i\}_{i=1}^{G} \sim \pi_\theta$, retained rubric set $\mathcal{R}_{\text{prev}}$ from the previous iteration (initialized to the weighted initial rubrics $\mathcal{R}_0 = \{(r_k, w_k)\}_{k=1}^{K}$ from Eq. 5 at the first iteration), rubric budget N_{new}, keep size M, generator–judge Φ

Ensure: rubric rewards $\{R_i^{\text{rub}}\}_{i=1}^{G}$; updated rubric set $\mathcal{K}$ carried to the next iteration

(i) Elicitation and judging

$\big(\mathcal{R}_{\text{new}}, \{j_{k,i}\}\big) \leftarrow \Phi\big(A, Q, y^\star, \{o_i\}_{i=1}^{G}, \mathcal{R}_{\text{prev}};\ N_{\text{new}}\big)$ ▷ $\leqslant N_{\text{new}}$ polarity-tagged new rubrics; verdicts $j_{k,i} \in \{0, 1\}$ for all rubrics

$\mathcal{R} \leftarrow \mathcal{R}_{\text{prev}} \cup \mathcal{R}_{\text{new}}$

for each rubric $r_k \in \mathcal{R}$ and each rollout $i = 1, \ldots, G$ **do**

  $b_{k,i} \leftarrow j_{k,i}$ **if** r_k is positive **else** $1 - j_{k,i}$ ▷ polarity normalization, Eq. 6

end for

(ii) Variance filtering

$s_k \leftarrow \mathrm{std}_i\{b_{k,i}\}_{i=1}^{G}$ for each $r_k \in \mathcal{R}$

$\mathcal{K} \leftarrow \{\, r_k \in \mathcal{R} :\ s_k > 0 \,\}$ ▷ drop criteria that all rollouts pass or all fail

if $\mathcal{K} = \varnothing$ **then**

  return $R_i^{\text{rub}} \leftarrow 0.5$ for all i ▷ neutral fallback: a constant reward contributes no advantage

end if

$\mathcal{K} \leftarrow$ top-M rubrics of $\mathcal{K}$ ranked by s_k ▷ keep the most discriminative criteria; all if fewer than M survive; $\mathcal{K}$ becomes $\mathcal{R}_{\text{prev}}$ next iteration

(iii) Weighting and scoring

$\{w_k\}_{k \in \mathcal{K}} \leftarrow \Phi\big(A, Q, \mathcal{K}\big)$ ▷ positive, $\sum_{k \in \mathcal{K}} w_k = 1$; supersedes the initial weights

return $R_i^{\text{rub}} \leftarrow \sum_{k \in \mathcal{K}} w_k\, b_{k,i}$ for $i = 1, \ldots, G$ ▷ Eq. 7; enters R_i in Eq. 9

rubric system and judging prompts corresponding to the first Φ call of Algorithm 1, and the weight assignment prompt corresponding to the second call.

Inference Prompt (MCQ Evaluation)

```
[System]
You are an expert audio understanding assistant. Listen carefully and answer
multiple-choice questions. Always think step by step inside <think> tags, then
give the final answer letter inside <answer> tags.

[User]  (audio attached as input_audio)
Listen to the audio carefully and answer the following multiple-choice question.

Question: {question}

Choices:
A. {choice_A}
B. {choice_B}
...

First, reason step by step inside <think> ... </think> tags.
Then output ONLY the letter (A, B, C, ...) of the correct answer inside
<answer> ... </answer> tags.

Example format:
<think>
Your reasoning here.
</think>
<answer>B</answer>
```

Figure 8: Inference prompt used for all benchmark evaluation (greedy decoding, temperature = 0).

Static Rubric Generation Prompt

```
You are an expert in audio understanding evaluation and rubric design.
Your task is to listen to a given audio clip and analyze the question-answer
pair, then generate exactly 5 evaluation rubrics that together assess response
quality for this question. The reference answer is only one possible response
from a student (not necessarily a good one).

# Input Data
[Audio]: <provided as audio attachment>
[Question]: {question}
[Reference Answer]: {response}

# The 5 rubric categories (use EACH category exactly once)
1. Auditory Evidence Identification & Grounding
2. Cross-Cue Verification & Distractor Elimination
3. Reasoning Clarity & Flow
4. Reasoning Focus & Efficiency
5. Domain-Specific Audio Techniques

# Rules
- Exactly 5 rubrics, ONE per category; no category repeats.
- Each rubric is binary (Yes/No) and MUST reference CONCRETE audio content
  actually heard in THIS clip or a specifically named distractor option.
- Each rubric carries a weight in (0,1); the 5 weights MUST sum to 1.0.
- DO NOT evaluate final-answer correctness (handled by a separate
  ground-truth reward).

# Output (JSON)
{
  "question_domain": "audio_reasoning/<sub-domain>",
  "rubrics": [{"category": ..., "criterion": ..., "weight": 0.xx} x5],
  "reference_score": x   # weights the reference would satisfy; must be < 0.5
}
```

Figure 9: Prompt for generating the initial (static) per-question rubrics.

Evolving Rubric System Prompt

```
You are an expert evaluator generating adaptive rubrics to assess model responses
on audio tasks, which take audio + text as input and must reason over what they hear.

## Task
Identify the most discriminative criteria that distinguish high-quality from
low-quality audio reasoning responses. Capture subtle quality differences that
existing rubrics miss particularly those unique to grounding answers in actual
acoustic evidence rather than text-only priors.

## Output Components
- Description: detailed, specific description of what makes a response excellent
  or problematic when reasoning over audio
- Title: concise abstract label (general, transferable across audio tasks)

## Categories
1. Positive Rubrics: excellence indicators distinguishing superior reasoning
2. Negative Rubrics: critical flaws definitively degrading reasoning quality

## Audio-Reasoning Quality Dimensions
- Acoustic Grounding: evidence actually present (words, sounds, prosody,
  timestamps) vs. plausible guesses; resistance to hallucinating audio content;
  faithfulness to what is audible vs. text/linguistic priors.
- Temporal & Sequential Reasoning: event ordering, durations/timestamps,
  tracking state changes across the timeline.
- Multi-Source Disentanglement: overlapping speakers, fore/background,
  speaker attribution/diarization, simultaneous events.
- Paralinguistic & Non-Lexical Reasoning: emotion, sarcasm, urgency, hesitation
  from prosody; what is said vs. how it is said; laughter/sighs/silence.
- Non-Speech Acoustic Understanding: sound-event ID, acoustic scene, music
  analysis (genre, tempo, key, instrumentation).
- Counting & Quantification: speakers, repetitions, distinct events, beats.
- Causal & Inferential Reasoning Over Audio: cues -> plausible causes; avoid
  spurious inference from a single ambiguous cue.
- Cross-Modal Integration: use the question to focus on relevant audio; do not
  let text priors override audio evidence when they conflict.

## Core Guidelines
1. Discriminative Power: only criteria that meaningfully separate the ACTUAL
   responses provided; exclude generic criteria ("is helpful").
2. Novelty & Non-Redundancy: never duplicate existing rubrics; add granular
   criteria if existing are broad; return empty lists if already comprehensive.
3. Avoid Mirror Rubrics: never create positive/negative versions of the same
   criterion; choose only the more discriminative direction.
4. Conservative Negative Rubrics: clear, observable failure modes (not mere
   absence of excellence); audio hallucination is a key negative axis.

## Selection Strategy
- Quantity: 1-5 total rubrics (fewer high-quality > many generic).
- More positive when responses are grounded but unsophisticated; more negative
  when systematic failures are present; empty when existing rubrics are comprehensive.

## Output Format (JSON)
{
  "question": "<original question verbatim>",
  "positive_rubrics": [{"description": ..., "title": ...}],
  "negative_rubrics": [{"description": ..., "title": ...}]
}

## Critical Reminders
- Each rubric must distinguish between the actual provided responses.
- Focus on objective, audio-grounded criteria a verifier could check against the audio.
- Treat audio-content hallucination and text-only reasoning that ignores the
  audio as first-class failure modes.
```

Figure 10: Full system prompt that drives per-step rubric evolution (Call 1).

Evolving Rubric Generation Judging Prompt

```
## Question
{question}

## Ground-truth final answer
{gt_letter}

## Existing rubrics
S1 ({category}): {criterion}
...
S5 ({category}): {criterion}

## Candidate Responses (N rollouts of the same prompt)
--- T1  [final answer: {ans} -> CORRECT / WRONG (chose X)] ---
{reasoning trace extracted from <think>...</think>}
--- TG ... ---

## Per-rollout correctness summary
T1: CORRECT ... TG: WRONG (chose B)
Note: a rollout's final-answer correctness is context only. Generate rubrics
that distinguish *reasoning quality*, not just correctness -- a rollout may be
CORRECT by guessing yet deserve a low score; a rollout may be WRONG yet show
good acoustic grounding.

## Required Output (strict JSON, no code fences)
Generate AT MOST {max_new} NEW rubrics non-redundant with S1..S5. Each new
rubric has a 'polarity': 'positive' (satisfied = good) or 'negative'
(satisfied = bad/flaw present). Then judge ALL rubrics (static + new) against
EACH of the 8 responses T1..T8 using exactly 'Yes' or 'No'.

Schema:
{"new_rubrics": [{"id":"N1","title":"<label>","description":"<criterion>",
                  "polarity":"positive|negative"}, ...],
 "judgments": {"S1":["Yes"|"No", x8], ..., "S5":[...], "N1":[...], ...}}
```

Figure 11: Call 1: jointly generate new candidate rubrics and judge all rubrics against the 8 rollouts.

Rubric Weight Assignment Prompt

```
You are an expert audio-reasoning evaluator. You will receive an audio clip, a
question, and a list of M binary rubrics that have been chosen as the most
discriminative for assessing model responses on this clip. Listen to the audio,
read the question, and decide how the M rubrics should be weighted: rubrics that
more directly probe whether the response is grounded in this specific audio
should receive higher weight. Output exactly M positive numbers that sum to 1.0.
Strict JSON output: {"weights": [w1, w2, ..., wK]}

## Question
{question}

## Kept Rubrics
R1: {rubric_text}
R2: {rubric_text}
...
RM: {rubric_text}

## Required Output
Return ONLY: {"weights": [..M floats summing to 1.0..]}
```

Figure 12: Call 2: assign normalized importance weights to the top-K kept rubrics.

Rubric Judging Prompt

```
You are an expert audio response evaluator. Listen to the attached audio, read
the question, candidate response, and 5 binary rubrics. For each rubric, decide
whether the candidate response satisfies it (Yes) or not (No).

[Question]: {question}

[Candidate Response]:
{response}

[Rubrics to evaluate]:
R1 ({category}, weight {w}): {criterion}
...
R5 ({category}, weight {w}): {criterion}

Return strictly JSON (no code fences):
{"R1": "Yes|No", "R2": "Yes|No", "R3": "Yes|No", "R4": "Yes|No", "R5": "Yes|No"}
```

Figure 13: Prompt for judging whether a response satisfies each rubric.