

# THINK SMART, NOT HARD: DIFFICULTY ADAPTIVE REASONING FOR LARGE AUDIO LANGUAGE MODELS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Large Audio Language Models (LALMs), powered by the chain-of-thought (CoT) paradigm, have shown remarkable reasoning capabilities. Intuitively, different problems often require varying depths of reasoning. While some methods can determine whether to reason for a given problem, they typically lack a fine-grained mechanism to modulate how much to reason. This often results in a “one-size-fits-all” reasoning depth, which generates redundant overthinking for simple questions while failing to allocate sufficient thought to complex ones. In this paper, we conduct an in-depth analysis of LALMs and find that an effective and efficient LALM should reason smartly by adapting its reasoning depth to the problem’s complexity. To achieve this, we propose a difficulty-adaptive reasoning method for LALMs. Specifically, we propose a reward function that dynamically links reasoning length to the model’s perceived problem difficulty. This reward encourages shorter, concise reasoning for easy tasks and more elaborate, in-depth reasoning for complex ones. Extensive experiments demonstrate that our method is both effective and efficient, simultaneously improving task performance and significantly reducing the average reasoning length. Further analysis on reasoning structure paradigm offers valuable insights for future work.

## 1 INTRODUCTION

In recent years, general artificial intelligence advances rapidly with the development of large language models (LLMs) (Dubey et al., 2024; Hurst et al., 2024; Team et al., 2024; Team, 2024). The reasoning ability of LLMs is further enhanced by the chain-of-thought (CoT) paradigm (Jaech et al., 2024; DeepSeek-AI et al., 2025; Yang et al., 2025), which significantly improves performance on complex problems. At the same time, Large Audio Language Models (LALMs) (Tang et al., 2024; Chu et al., 2023; 2024) also progress rapidly and raise an important question: how can they achieve reasoning more efficiently and effectively?

In previous work, Xie et al. (2025) enable reasoning in LALMs through supervised fine-tuning (SFT) on a large-scale dataset with CoT annotations, providing the first evidence that CoT is effective for solving complex audio understanding problems. Building on this, Li et al. (2025) introduces group relative policy optimization (GRPO) into LALMs for the first time, surpassing the former with less data and offering initial comparisons across different prompts. These studies focus either on SFT or GRPO, and both apply reasoning to all questions, but they lack in-depth analysis of the differences between the two approaches. Therefore, we conduct a detailed analysis to determine under what conditions SFT and GRPO are more effective. We find that GRPO performs better on harder questions, whereas on easier ones it tends to produce redundant reasoning and performs slightly worse than SFT. Furthermore, we analyze GRPO under forced reasoning (explicit prompt) and without forced reasoning (implicit prompt), and observe that the forced reasoning models maintain a clear advantage on harder questions, while the two settings perform similarly on easier ones. Taken together, these findings indicate that achieving efficient reasoning in LALMs requires adapting reasoning length to problem difficulty.

Regarding reasoning length and efficiency, Qu et al. (2025) highlight issues such as redundancy and overthinking. RL-based studies on LLMs (Arora & Zanette, 2025; Aggarwal & Welleck, 2025) design length-penalty rewards but rely on fixed thresholds that overlook question types and difficulty levels. In the audio domain, Wu et al. (2025) introduces a “when to think” mechanism that guides the

model on whether reasoning is necessary. However, for the samples that still perform reasoning, it lacks a fine-grained mechanism to modulate how much to reason. Therefore, an effective and efficient LALM should reason smartly by adapting its reasoning depth to the problem’s complexity, achieving short reasoning for simple questions and deeper reasoning for difficult ones. Based on this analysis and prior work, we propose a new length-based reward function that no longer depends on fixed thresholds. In addition, we introduce two difficulty-adaptive standards to complement this reward, enabling reasoning length to align more appropriately with question difficulty. Both approaches achieve strong performance on the MMAU benchmark (Sakshi et al., 2025), particularly on harder questions, while also producing much shorter reasoning than direct GRPO models, thereby greatly improving reasoning efficiency. Furthermore, we conduct a qualitative case study on models from the main experiments, providing a detailed analysis of the reasoning structures in their outputs.

In conclusion, the main contributions are as follows:

- In this paper, we conduct in-depth analyses of LALMs and show that a smart LALM should reason adaptively, adjusting its reasoning depth to match the complexity of the problem.
- Based on this analysis, we propose two difficulty-adaptive length-based rewards, aiming to shorten reasoning length while achieving concise reasoning for simple questions and encouraging deeper reasoning for difficult ones.
- We carry out extensive experiments on the two proposed rewards, both of which achieve strong performance on the MMAU benchmark while significantly reducing reasoning length. In addition, we provide a qualitative analysis of output paradigms across models, offering useful guidance for future work.

We have provided an anonymous GitHub repository <https://anonymous.4open.science/r/ICLR2026-Anonymous-Repo-24196/> to support the reproduction of this work and will release all our experimental models after the double-blind review process.

## 2 DEEP ANALYSIS OF DIFFERENT METHODS FOR LALMS

LALMs have already demonstrated strong capability in addressing basic understanding tasks. Recent studies (Xie et al., 2025; Li et al., 2025) focus on enhancing their ability to solve complex problems through reasoning. However, while these works adopt different implementation methods, they lack in-depth analysis of the differences between approaches. Therefore, in this section, we conduct detailed analysis from two perspectives: which is more effective under various conditions between SFT and GRPO, and whether performance gains are driven by explicit or implicit prompt.

**Data.** For training, we use two datasets, FS and AVQA. FS is constructed on the basis of AVQA by incorporating four additional datasets, covering three different task types. Its size is approximately twice that of AVQA, and the detailed distribution is presented in Table 1.

Table 1: The data distribution of the FS training set, including different task types and their corresponding data sources.

Task	Dataset-Source	Num
Audio Grounding	AudioGrounding (Xu et al., 2021)	1,805
Sound Classification	VocalSound (Gong et al., 2022)	15,531
	TUT2017 (Mesaros et al., 2017)	3,744
Sound Question Answering	Clotho-AQA (Lipping et al., 2022)	6,615
	AVQA (Yang et al., 2022)	36,036

**Setup.** For the base models, we primarily adopt Qwen2-Audio-7B-Instruct<sup>1</sup> and Qwen2.5-Omni-7B<sup>2</sup>. More detailed experimental settings, including the two prompts, are provided in Appendix

<sup>1</sup><https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct>

<sup>2</sup><https://huggingface.co/Qwen/Qwen2.5-Omni-7B>

A.2. At the same time, for those experiments that share common settings, we also add an additional group average in Table 2 to facilitate comparisons across groups.

## 2.1 HOW DOES GRPO REASONING DIFFER FROM DIRECT ANSWERS WITH SFT?

Here we mainly compare the “SFT on Qwen2-Audio-7B-Instruct” part and the “GRPO on Qwen2-Audio-7B-Instruct” part of Table 2. From the “Average” of the SFT part and the “Average” of the Prompt2 part, we observe that SFT performs very well on easy-level questions, while GRPO shows greater advantages on medium and hard questions. We believe this is mainly because, when facing medium and harder questions, the model cannot rely only on the base knowledge learned during pretraining to give direct answers and instead needs to learn how to use this knowledge through reasoning. In contrast, its weaker performance on easy questions is largely due to redundant reasoning content or errors made during the reasoning process that are carried over into the final answer, leading to incorrect results.

Therefore, for the question in this subsection, we conclude that GRPO is more effective on complex tasks that models cannot solve directly, but its reasoning on simple ones still needs further optimization to reduce redundancy and potential error propagation.

Table 2: The performance of different models under different training paradigms, fine-tuning strategies, training datasets, and prompting styles.

Models	Airbench-Foundation		MMAU-Test-Mini					
	Sound	Sound	Music	Speech	Easy	Medium	Hard	Avg
SFT On Qwen2-Audio-7B-Instruct								
On FS, Full	80.49	59.76	58.38	62.46	57.36	64.31	54.31	60.20
On AVQA, Full	67.06	60.36	56.59	59.46	53.10	63.92	53.88	58.80
On FS, LoRA	77.74	68.17	65.57	64.86	59.30	74.51	55.60	66.20
On AVQA, LoRA	67.38	66.07	60.48	54.65	49.22	69.80	52.16	60.40
Average	73.16	63.59	60.25	60.35	54.74	68.13	53.96	61.40
GRPO On Qwen2-Audio-7B-Instruct								
On FS, Full, Prompt2	81.10	67.57	64.67	62.76	55.04	73.73	56.90	65.00
On AVQA, Full, Prompt2	70.35	67.87	66.77	60.96	52.33	75.10	57.76	65.20
On FS, LoRA, Prompt2	69.60	69.37	60.48	55.26	49.61	71.76	53.02	61.70
On AVQA, LoRA, Prompt2	69.38	66.97	59.28	56.16	47.29	70.78	53.88	60.80
Average	72.61	67.95	62.80	58.79	51.07	72.84	55.39	63.18
On FS, LoRA, Prompt1	69.87	67.27	61.98	56.76	50.00	71.57	54.31	62.00
On AVQA, LoRA, Prompt1	68.10	64.86	60.18	53.15	47.29	71.18	50.43	60.20
Average	68.99	66.07	61.08	54.96	48.65	71.38	52.37	61.10
GRPO On Qwen2.5-Omni-7B								
On FS, Full, Prompt2	<b>83.46</b>	72.37	<b>67.66</b>	68.76	59.30	<b>78.43</b>	<b>61.63</b>	<b>69.60</b>
On FS, LoRA, Prompt2	76.86	<b>73.57</b>	65.56	<b>69.06</b>	<b>59.69</b>	78.23	60.77	69.40
Average	80.16	72.97	66.61	68.91	59.49	78.33	61.20	69.50

## 2.2 DOES PERFORMANCE COME FROM EXPLICIT REASONING OR IMPLICIT ACTIVATION?

This part mainly compares the four LoRA experiments under the “GRPO on Qwen2-Audio-7B-Instruct” part. First, it should be noted that the two prompts produce outputs with clear differences. Models trained with the implicit prompt do not generate a reasoning process for nearly every sample as those trained with the explicit prompt do; instead, they often directly produce answers in a way similar to SFT. When comparing results within the “Prompt1” (implicit prompt) experiments, models trained on the larger FS dataset consistently achieve better performance. These results suggest that using implicit prompts introduces SFT-like characteristics, relying on larger datasets to achieve stronger generalization, whereas explicit reasoning allows the model to truly learn from the data, which in turn demonstrates the necessity of explicit reasoning. When further comparing “Prompt1”

and “Prompt2” (explicit prompt), the performance gap is generally small, with “Prompt2” outperforming “Prompt1” by about 0.15 on average. Looking at the details, “Prompt1” performs about 0.1 better on easy and medium questions, but lags behind by 1.1 on hard questions. These results further show that reasoning on easier questions can lead to redundancy and error propagation, while harder questions require deeper reasoning.

Overall, for the question in this subsection, we conclude that reasoning for all questions or not reasoning at all is not the optimal solution. Instead, the model should learn to adjust reasoning length according to different questions—reducing reasoning length for those that do not require it, while increasing reasoning for those that lack it. In this way, it can achieve both performance improvements and efficiency gains.

### 3 ENHANCING LALMS WITH DIFFICULTY-AWARE ADAPTIVE REASONING

The above results and analysis indicate that different types of questions require different reasoning lengths. Therefore, we aim to link question difficulty with reasoning length, enabling shorter reasoning for simple questions and deeper reasoning for difficult ones. Specifically, we first define two model-perspective difficulty-adaptive standards: one based on group accuracy of rollout samples and the other based on the audio attention of the current sample. We then apply these difficulty standards to a rule-based reward function that varies with reasoning length, thereby linking question difficulty with reasoning length. In the following, we elaborate on these two components in detail.

#### 3.1 DEFINING MODEL-PERSPECTIVE DIFFICULTY

As mentioned above, our core idea is to encourage models to reason more on difficult questions and less on simple ones. Thus, our main approach is to dynamically adjust the reward based on both question difficulty and reasoning length. At the same time, since there are gaps between different definitions of difficulty perspectives, the model’s own perspective during training can better reflect the actual situation. Therefore, in this subsection, we provide a detailed explanation of the two proposed model-perspective difficulty-adaptive standards. We refer to these as **Group Ratio Difficulty Reward (GRDR)** and **Group Audio Attention Difficulty Reward (GA<sup>2</sup>DR)**.

**GRDR.** The first is based on the ratio of correct samples within a rollout group. For example, when the group size  $G = 8$ , if six or more responses are correct, the question is labeled as easy; if fewer than six but at least three are correct, it is labeled as medium; and if fewer than three are correct, it is labeled as hard.  $\gamma$  is used to represent the difficulty value of a question, where larger values correspond to more difficult questions. The corresponding  $\gamma$  values are 0, 0.5, and 1, respectively, with the specific formula as follows:

$$\gamma = \begin{cases} 0, & C \geq 6, \\ 0.5, & 3 \leq C < 6, \quad (G = 8), \quad C = \sum_{i=1}^G c_i \\ 1, & C < 3, \end{cases} \quad (1)$$

Here,  $c_i$  indicates whether the answer of the rollout sample  $o_i$  is correct, taking values of either 0 or 1.  $C$  correspondingly represents the number of correct answers within a rollout group.

**GA<sup>2</sup>DR.** The second approach is more characteristic of the audio modality. As mentioned earlier, when the model’s attention to the audio segment is more dispersed, it suggests that the audio is complex and that the model struggles to identify a key point to solve the problem. In this case, the attention entropy is relatively large, corresponding to higher difficulty. Specifically, we use  $a_j^{(n)}$ , the attention after softmax from the last token position in the final hidden layer over all previous positions, and  $\bar{p}_j$ , the attention values assigned to audio tokens averaged across attention heads  $N$ , and then compute the entropy. This entropy is normalized across the batch to a value in  $[0, 1]$ , which directly represents difficulty. The complete calculation process is as follows:

$$a_j^{(n)} = \mathbf{A}_{T,j}^{(n)}, \quad \bar{p}_j = \frac{1}{N} \sum_{n=1}^N a_j^{(n)}, \quad H = - \sum \bar{p}_j \log \bar{p}_j, \quad (j \in \mathcal{M}). \quad (2)$$

$$\gamma^{(b)} = \frac{H^{(b)} - \min_{b' \in \mathcal{B}} H^{(b')}}{\max_{b' \in \mathcal{B}} H^{(b')} - \min_{b' \in \mathcal{B}} H^{(b')}} \in [0, 1]. \quad (3)$$

Here,  $\mathbf{A}$  denotes the complete attention matrix,  $T$  represents the number of tokens,  $\mathcal{M}$  indicates the indices corresponding to the audio attention part,  $H$  denotes the computed entropy, and  $\mathcal{B}$  is the batch size. The final  $\gamma^{(b)}$  represents the difficulty value of the  $b$ -th sample within a batch. Compared with the first method, this approach provides a more fine-grained division of difficulty, no longer limited to three categories, but instead yielding a continuous difficulty range.

### 3.2 DIFFICULTY-ADAPTIVE LENGTH-BASED REWARD

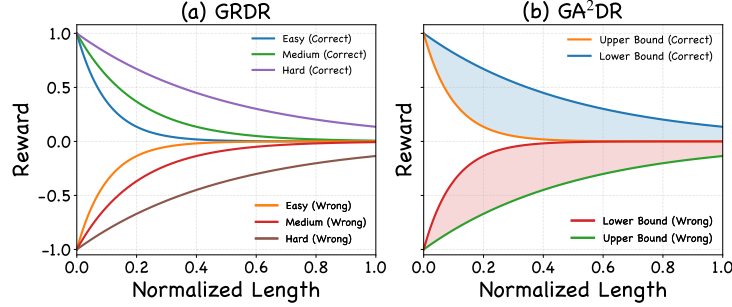


Figure 1: Curves of GRDR and GA²DR with normalized length.

After defining question difficulty, the next step is to link difficulty with reasoning length in an effective way. Our core idea is short reasoning for simple questions and long reasoning for difficult ones, with reward values that change dynamically rather than staying fixed. Thus, we design a rule-based reward using a negative exponential function, with the corresponding curve shown in Figure 1. Specifically, it smoothly adjusts reward values according to the ratio between the current reasoning length and the model’s maximum reasoning length. The decision between reward and penalty also depends on correctness. Incorrect samples always receive a penalty that decreases as reasoning length grows, encouraging further reasoning. Once a sample is answered correctly at a certain stage, the penalty switches to a positive reward, with the reward value increasing as reasoning length shortens, encouraging concise reasoning that retains only the core solution steps. For difficulty, specifically in GRDR (plot a), the difficulty levels are divided into three categories corresponding to three curves. Along the correct reward axis, difficulty increases from bottom to top, while on the wrong axis, it decreases accordingly. In contrast, GA²DR (plot b) does not fix discrete difficulty levels but allows the curve exponent to vary continuously within the region. Both follow the same trend in which simple questions correspond to steeper curves and difficult ones to flatter curves, meaning that under the same reward value, difficult questions lead to longer reasoning lengths, and under the same reasoning length, simple questions yield smaller rewards. This achieves short reasoning for simple tasks and longer reasoning for difficult ones. The detailed calculation is as follows:

$$r_i = \text{sign}(o_i) \cdot e^{-k(\gamma)l_{o_i}} \quad (4)$$

$$k(\gamma) = (1 - \gamma)k_{\text{easy}} + \gamma k_{\text{hard}} \quad (5)$$

Here, the *sign* function indicates whether a sample is correct or incorrect, taking only values 1 or -1.  $l_{o_i}$  is the ratio of the sample’s output length to the model’s maximum output length, ranging from [0,1].  $k(\gamma)$  is obtained by applying linear interpolation to map the difficulty standards defined in the first part onto optimization curves with different slopes.

## 4 EXPERIMENTS

**Data.** For training, we use FS in this section, which is approximately twice the size of AVQA. The detailed data distribution has been presented in Section 2.

**Setup.** In this section, our experiments are mainly based on Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B. All models are fine-tuned using LoRA with Prompt2. Most evaluations are conducted on MMAU-test-mini with ACC as the metric. And we also evaluate our proposed methods on MMAU-test-mini (v05.15.25), which improves the Q&A formulation and enhances the quality of the audio itself compared to the previous version, on MMAR (Ma et al., 2025), a benchmark designed to assess the deep reasoning capabilities of Audio Language Models (ALMs) in complex settings that span large-scale multitask, multimodal, and multilingual scenarios. Other details can be found in Appendix A.2.

Table 3: The performance of models on MMAU-test-mini with different base models and reward settings, reported under both the human-perspective difficulty annotations and the assigned model-perspective difficulty annotations. Here,  $\dagger$  denotes the model-perspective difficulty annotations and \* denotes a leading proprietary model.

Models	MMAU-Test-Mini									
	Sound	Music	Speech	Easy	Easy $\dagger$	Medium	Medium $\dagger$	Hard	Hard $\dagger$	Avg
Baseline Models										
Qwen2-Audio-7B-Instruct	53.75	48.80	47.74	48.06	62.80	50.58	45.32	51.29	28.18	50.10
Qwen2.5-Omni-7B	67.26	59.88	53.75	54.26	90.89	71.96	44.39	41.37	11.19	60.30
Kimi-Audio-7B-Instruct	72.37	58.98	61.66	50.38	91.84	75.88	52.80	54.74	18.14	64.40
Gemini2.5-Pro-0506* <sup>1</sup>	70.57	65.26	62.16	52.32	95.82	77.64	57.47	55.60	12.35	66.00
Based On Qwen2-Audio-7B-Instruct										
GRPO	69.37	60.48	55.26	49.61	82.82	71.76	47.66	53.02	30.16	61.70
+ TR	68.16	60.77	55.85	48.83	83.87	71.56	46.26	53.87	28.95	61.60
+ GRDR	66.96	58.38	60.06	54.26	81.59	69.60	45.32	53.01	35.13	61.80
Based On Qwen2.5-Omni-7B										
GRPO	<b>73.57</b>	65.56	<b>69.06</b>	59.69	<b>93.92</b>	78.23	59.81	<b>60.77</b>	27.41	69.40
+ TR	72.97	66.46	65.16	58.14	93.73	78.43	57.47	57.75	25.86	68.40
+ GRDR	71.47	<b>72.45</b>	66.66	<b>60.07</b>	93.16	80.00	58.87	59.91	<b>32.81</b>	<b>70.20</b>
+ GA <sup>2</sup> DR	71.47	71.85	66.66	57.75	92.78	<b>80.58</b>	<b>59.81</b>	60.34	32.04	70.00

Table 4: Evaluations on MMAU-v0515. Here, \* denotes a leading proprietary model.

Models	MMAU-Test-Mini-v0515			
	Sound	Music	Speech	Avg
Baseline Models				
Qwen2-Audio	62.16	62.27	55.55	60.00
Qwen2.5-Omni	74.17	65.26	61.56	67.00
Kimi-Audio	78.97	60.47	66.96	68.80
Gemini2.5-Pro*	76.57	73.95	80.78	77.10
Based On Qwen2.5-Omni-7B				
GRPO	84.08	69.46	74.17	75.90
+ TR	83.78	70.65	74.47	76.30
+ GRDR	83.48	70.35	75.97	76.60
+ GA <sup>2</sup> DR	83.18	71.55	75.67	<b>76.80</b>

Table 5: Evaluations on MMAR. Here, \* denotes a leading proprietary model.

Models	MMAR			
	Sound	Music	Speech	Avg
Baseline Models				
Qwen2-Audio	33.33	24.27	32.31	30.00
Qwen2.5-Omni	58.79	40.78	59.86	56.70
Kimi-Audio	57.57	45.63	63.26	59.00
Gemini2.5-Pro*	73.33	64.07	88.77	80.50
Based On Qwen2.5-Omni-7B				
GRPO	60.00	48.05	62.24	59.90
+ TR	64.84	49.51	63.94	61.90
+ GRDR	61.21	51.94	65.30	61.20
+ GA <sup>2</sup> DR	64.84	54.85	65.30	<b>62.90</b>

#### 4.1 MAIN RESULTS

Table 3 mainly presents three parts: the performance of four baseline models and our proposed rewards on Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B. We further report results on MMAU-

<sup>1</sup><https://ai.google.dev/gemini-api/docs/models#gemini-2.5-pro>

v0515 and MMAR, together with the base model on these benchmarks, shown in Table 4 and Table 5. TR denotes the basic Truncation Reward (Liu et al., 2025), with its formula provided in Appendix A.2. In addition, we extend MMAU-test-mini with model-perspective difficulty labels annotated by the four baseline models in Table 3, with details given in Appendix A.3.

**Performance across Models.** First, we analyze the four baseline models. Their overall performance reflects different capability levels and the typical range of most systems. However, they differ markedly on hard questions. Even Gemini2.5-Pro, which is nearly the strongest on average, performs poorly on hard items, while Qwen2-Audio, the weakest on average, shows a clear advantage. We attribute this to the combined effects of the LLM backbone strength and the number of supported modalities. For easy questions, the LLM backbone is dominant because these items require minimal audio understanding and rely mainly on text comprehension. In contrast, hard questions demand stronger audio interpretation, where pure LALMs often outperform Omni-style models that integrate more modalities. This pattern is evident in Qwen2.5-Omni and Kimi-Audio, which share a “Whisper + Qwen2.5” architecture but still exhibit a noticeable gap on hard questions.

Second, we compare our proposed GRDR and GA<sup>2</sup>DR with TR. For Qwen2-Audio-7B-Instruct, the three methods achieve similar overall performance, with GRDR performing best and TR worst, confirming the effectiveness of our approach. TR shows slightly better performance on medium questions, but GRDR is clearly stronger on hard ones. For Qwen2.5-Omni-7B, both of our methods deliver clear overall gains. They substantially outperform TR on medium and hard questions, while maintaining comparable results on easy ones. This shows that our methods effectively utilize question difficulty to assign appropriate rewards, promote deeper reasoning on hard questions, and improve performance on challenging tasks. Unlike TR, which uses a fixed truncation length and only constrains part of the samples, our reward designs offer more balanced treatment across difficulty levels, leading to stronger overall results.

Third, we compare GRDR and GA<sup>2</sup>DR, focusing on the results based on Qwen2.5-Omni-7B. The two methods differ in that GRDR is outcome-oriented, whereas GA<sup>2</sup>DR is process-oriented. GRDR performs better on easy questions, GA<sup>2</sup>DR is stronger on medium ones, and their performance is similar on hard questions. We believe this is due to their different difficulty definitions: GRDR uses only three levels, while GA<sup>2</sup>DR applies a finer, unconstrained difficulty scale, which particularly benefits medium questions. For easy and hard questions, GA<sup>2</sup>DR is slightly weaker because normalization may place samples with similar audio attention entropy into different difficulty bins, diminishing its advantage.

In summary, both GRDR and GA<sup>2</sup>DR achieve clear performance gains, especially on hard questions, showing that our methods can effectively adapt to different questions according to their difficulty.

**Performance across Benchmarks.** To further evaluate our model in broader and more reasoning-intensive scenarios, we conduct additional evaluations on MMAU-v0515 and MMAR, which also help verify the generalization ability of our methods. In terms of overall performance, our two methods still achieve the best results on MMAU-v0515 compared with all baselines. On MMAR, however, the two methods behave differently: GA<sup>2</sup>DR maintains a clear performance lead, whereas GRDR falls behind the TR method.

This difference can be explained by the nature of the two approaches. GRDR is outcome-oriented, while GA<sup>2</sup>DR is process-oriented. GRDR is highly susceptible to noise in rollout samples, making it prone to reward hacking. As a result, it performs reasonably well on the relatively simpler MMAU benchmarks but degrades significantly on the more challenging MMAR benchmark. In contrast, GA<sup>2</sup>DR is unaffected by such noise because it determines difficulty solely based on the model’s current behavior and the audio characteristics of each question. This removes much of the randomness and reduces the likelihood of reward hacking, leading to more stable performance across benchmarks and consistently strong results on more challenging tasks. We also conduct additional experiments to further verify generalization. The results confirm that GRDR is indeed affected by random noise introduced during rollout; however, once this issue is mitigated through appropriate constraints, its performance still surpasses all baseline methods. Detailed results and analysis are provided in Appendix A.4.

## 4.2 ANALYSIS OF REASONING LENGTH ACROSS MODELS

In this subsection, we mainly compare reasoning length from both difficulty perspectives, as shown in Figure 2, focusing on the four baseline models and all models based on Qwen2.5-Omni-7B. Because the length gaps between models are relatively large, we present log-scaled lengths in this part. We also provide Figure 3, which shows three Qwen2.5-Omni-7B-based models—direct GRPO and our two proposed rewards. To better illustrate how our methods adjust reasoning length across different difficulty levels, this figure uses the actual token counts without log-scaling. In addition, detailed length statistics are provided in Appendix A.5.

First, examining length trends under the two perspectives of difficulty, clear differences emerge. Under the human perspective, reasoning is longer for easy and hard questions, with medium questions shortest, whereas under the model perspective, reasoning length increases with difficulty. This indicates a fundamental inconsistency between human and model standards. Consequently, using human-perspective difficulty in training may conflict with the model’s perspective. Furthermore, since different models within the same perspective already show varying trends, models at different training stages naturally behave like distinct models, each with its own patterns. Together, these observations strongly support adopting model-perspective difficulty in training and continuously updating the difficulty standard as the model improves.

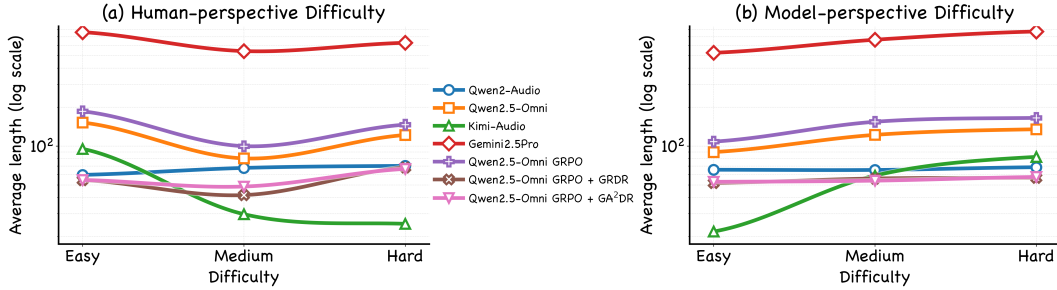


Figure 2: The trend of average length across different models on MMAU-Test-Mini, under both the human-perspective difficulty and model-perspective difficulty. The length is measured in tokens and is presented after applying a logarithmic transformation.

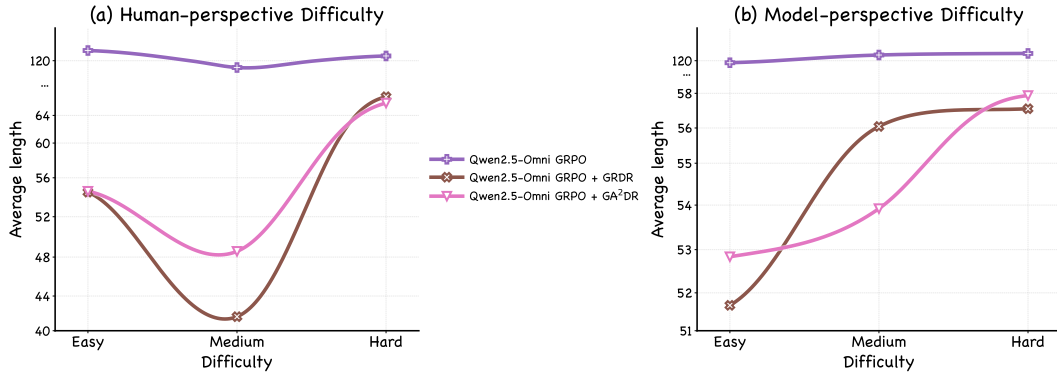


Figure 3: The trend of average reasoning length for direct GRPO and our two proposed methods on MMAU-Test-Mini, evaluated under both human-perspective and model-perspective difficulty. Length is measured directly in tokens without any logarithmic transformation.

Second, when comparing reasoning lengths of our proposed rewards with other models, we find that our two rewards produce shorter reasoning across all difficulty levels than most other models. Under the human perspective, the lengths of our methods are close to direct GRPO for easy and hard questions but are much shorter on medium questions. Under the model perspective, the curves of our two rewards almost overlap and show clear improvements over both the base model and direct GRPO at all difficulty levels. Moreover, under the model perspective, reasoning length increases



appropriately with difficulty. These results strongly demonstrate the effectiveness of our proposed rewards, achieving shorter reasoning for easy questions and deeper reasoning for hard ones.

In summary, our proposed difficulty-adaptive length-based reward is both reasonable and effective. Specifically, it achieves short reasoning for simple questions and long reasoning for difficult ones, while overall significantly reducing reasoning length. This makes LALMs reason smartly, achieving greatly improved reasoning efficiency alongside higher performance.

#### 4.3 ABLATION STUDY: RESULTS OF SINGLE LENGTH-BASED REWARD AND $k$ SETTINGS

In this section, we conduct an in-depth ablation study on our proposed difficulty-adaptive length-based reward. Specifically, we remove the mechanism in Equation 5 that determines different reward curves based on question difficulty, and replace  $k(\lambda)$  in Equation 4 with a fixed value  $k$ , meaning that all questions share the same reward curve. We also experiment with different settings of  $k$  to further validate the effectiveness of our approach. The main evaluation is performed on MMAU-Test-Mini in Table 6, and additional results on MMAU-v0515 and MMAR are also provided in Table 7 and Table 8.

Table 6: The performance of models trained with the single length-based reward setting. Here,  $\dagger$  denotes the model-perspective difficulty annotations.

Models	MMAU-Test-Mini									
	Sound	Music	Speech	Easy	Easy $^\dagger$	Medium	Medium $^\dagger$	Hard	Hard $^\dagger$	Avg
GRPO Based On Qwen2.5-Omni-7B (With Length-based Reward Only)										
k=2	73.27	67.06	67.86	59.69	94.49	77.84	58.41	61.63	27.41	69.40
k=6	72.37	66.46	68.76	56.20	93.92	79.21	57.00	61.63	28.95	69.20
k=10	72.37	70.06	66.36	58.91	93.92	79.60	56.07	59.48	31.27	69.60

Table 7: Evaluations on MMAU-v0515.

Models	MMAU-Test-Mini-v0515			
	Sound	Music	Speech	Avg
GRPO Based On Qwen2.5-Omni-7B				
k=2	81.38	73.05	74.77	76.40
k=6	82.88	71.55	74.77	76.40
k=10	81.98	71.25	76.57	76.60

Table 8: Evaluations on MMAR.

Models	MMAR			
	Sound	Music	Speech	Avg
GRPO Based On Qwen2.5-Omni-7B				
k=2	61.81	50.97	63.26	61.40
k=6	60.60	52.91	64.96	61.80
k=10	61.81	52.91	64.62	62.20

From the perspective of overall performance, all fixed  $k$ -value settings outperform direct GRPO and the traditional TR method. This demonstrates the effectiveness of our length-based reward and shows that these samples indeed require length optimization—direct GRPO tends to produce either redundant or insufficient reasoning. It also indicates that, compared with TR, our reward design can dynamically adjust rewards based on length rather than relying on a fixed threshold, resulting in better optimization. Comparing with the difficulty-adaptive methods GRDR and GA<sup>2</sup>DR, most fixed  $k$  settings perform worse, further confirming the validity and effectiveness of incorporating adaptive difficulty. Across the two MMAU benchmarks, performance fluctuates noticeably across all three difficulty levels under different  $k$  values, showing that applying a single reward curve to all questions leads to imbalance and reduced overall performance. Overall, these experiments further validate the effectiveness of our difficulty-adaptive length-based reward.

## 5 RELATED WORK

In recent years, advances in LLMs have driven the development of MLLMs, enhancing multimodal understanding. In the audio domain, LALMs such as Qwen2-Audio (Chu et al., 2024), Audio Flamingo (Kong et al., 2024), and SALMONN (Tang et al., 2024) handle basic understanding well

but remain limited on complex tasks due to short outputs and lack of reasoning. Later models like Qwen2.5-Omni (Xu et al., 2025) and Kimi-Audio (KimiTeam et al., 2025) demonstrate some initial reasoning ability but still rely heavily on SFT, leaving outputs fixed and dependent on pretraining data, thus performing poorly in complex scenarios. To address this, some studies extend SFT, such as Audio-Reasoner (Xie et al., 2025), which uses large-scale CoT-annotated pairs to achieve reasoning via SFT. Others adopt RL, as in R1-AQA (Li et al., 2025) and Omni-R1 (Rouditchenko et al., 2025), which avoids using data with CoT, and promotes self-driven reasoning.

However, prior work mainly focuses on overall performance, leaving open key questions: how GRPO differs from SFT, and whether improvements come from explicit reasoning or from prompts activating implicit reasoning. Therefore, in this study, we conduct a systematic and deep analysis of these two questions to provide a clearer understanding of different approaches in LALMs. Based on this in-depth analysis, we draw an important conclusion: The explicit reasoning introduced by GRPO is necessary, especially for more difficult questions, but it often leads to redundant reasoning on simple ones. Therefore, reasoning length should be optimized according to problem difficulty, reducing redundancy for simple questions while encouraging deeper reasoning for harder ones.

Reasoning efficiency thus becomes a major challenge. Prior studies (Qu et al., 2025) highlight inefficiencies such as redundant content and overthinking. RL-based methods explore reward function designs to address this issue (Arora & Zanette, 2025; Aggarwal & Welleck, 2025; Shen et al., 2025). Most methods rely on setting a fixed length threshold for optimization. On the one hand, such approaches do not link reasoning length with problem type and therefore cannot adapt to all kinds of problems. On the other hand, the fixed threshold cannot change with the model’s evolving ability during training. Liu et al. (2025) summarizes these efforts and proposes a difficulty-aware dynamic approach that improves both performance and efficiency. However, this method depends on an additional independent dataset during training to measure the model’s capability, which not only increases computational cost but also makes the results sensitive to how this dataset is selected. In LALMs, Audio-Thinker (Wu et al., 2025) instead tackles the problem through a “when to think” mechanism, dividing tasks into those requiring reasoning and those that do not, but it does not further optimize the samples that require reasoning. In this work, we propose two difficulty-adaptive length-based rewards to enable efficient and effective reasoning in LALMs. Our method maintains comparable or even superior performance while reducing overall reasoning length, encouraging concise reasoning for simple questions, deeper reasoning for difficult ones, and progressive optimization as the model’s capability improves.

## 6 LIMITATIONS AND FUTURE WORK

In this work, we optimize reasoning across difficulty levels and achieve notable gains. But, limitations still remain. Our analysis of outputs from all models—including prior work and proprietary models—indicates that a strong CoT response should extract key information, perform structured reasoning, and deliver a clear final answer. Detailed analyses and examples are provided in Appendix A.9 and Appendix A.10. We also believe that incorporating rewards from external LLM APIs may further enhance the coherence and reliability of CoT outputs while preserving the intended reasoning paradigm, which we will explore in future work.

## 7 CONCLUSION

In this work, we focus on addressing the question of how LALMs can achieve reasoning more efficiently and effectively. First, we conduct detailed experimental analyses on two key issues: in which conditions SFT and GRPO are more effective, and whether performance improvements come from explicit prompts directly or from implicit prompts activating the model’s reasoning ability. Our findings show that explicit reasoning through GRPO is more effective, but the reasoning length should be optimized according to question difficulty. Based on this, we propose two difficulty-adaptive length-based rewards and carry out extensive experiments. The results demonstrate that our method achieves better overall performance, significantly shortens reasoning length, and improves efficiency. We also provide qualitative analyses of reasoning outputs from multiple models, identify an ideal reasoning structure paradigm, and recommend a set of training procedures for models with different capability levels, offering useful guidance for future work.

## ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation is involved. All datasets used are sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We take care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information is used, and no experiments are conducted that raise privacy or security concerns. We remain committed to maintaining transparency and integrity throughout the research process.

## REPRODUCIBILITY STATEMENT

In this section, we provide a detailed description of the reproduction details of this work.

Our base models, training framework, and datasets are all publicly available, with detailed information provided in Appendix A.2. In addition, we provide an anonymous GitHub repository <https://anonymous.4open.science/r/ICLR2026-Anonymous-Repo-24196/> to support the reproduction of this work. The repository contains scripts for the two proposed difficulty-adaptive length-based rewards in this paper, as well as the baseline TR method, which can be directly integrated into the ms-swift framework. It also includes all evaluation scripts and reasoning-length assessment scripts, the Cold-Start dataset used in our experiments (including both audio wav files and JSONL annotations), and the outputs of different models on the 20 sarcasm cause detection questions discussed in the analysis.

## REFERENCES

- Pranjal Aggarwal and Sean Welleck. L1: controlling how long A reasoning model thinks with reinforcement learning. *CoRR*, abs/2503.04697, 2025.
- Daman Arora and Andrea Zanette. Training language models to reason efficiently. *CoRR*, abs/2502.04463, 2025.
- 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.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *CoRR*, abs/2407.10759, 2024.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv-2407, 2024.

- Yuan Gong, Jin Yu, and James R. Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichen, Ian O’Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya. Openai o1 system card. *CoRR*, abs/2412.16720, 2024.
- KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report. *CoRR*, abs/2504.18425, 2025.
- Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*. ACM, 2023.
- 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. *CoRR*, abs/2503.11197, 2025.
- Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowdsourced dataset for audio question answering. In *30th European Signal Processing Conference, EUSIPCO 2022, Belgrade, Serbia, August 29 - Sept. 2, 2022*. IEEE, 2022.
- Wei Liu, Ruochen Zhou, Yiyun Deng, Yuzhen Huang, Junteng Liu, Yuntian Deng, Yizhe Zhang, and Junxian He. Learn to reason efficiently with adaptive length-based reward shaping. *CoRR*, abs/2505.15612, 2025.
- Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Chng Eng Siong, and Xie Chen. MMAR:

- A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *CoRR*, abs/2505.13032, 2025.
- Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. DCASE2017 challenge setup: Tasks, datasets and baseline system. In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2017, Munich, Germany, November 16-17, 2017*, 2017.
- Ruotian Peng, Yi Ren, Zhouliang Yu, Weiyang Liu, and Yandong Wen. Simko: Simple pass@k policy optimization. *arXiv preprint arXiv:2510.14807*, 2025.
- Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *CoRR*, abs/2503.21614, 2025.
- Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogério Feris, and James R. Glass. Omni-r1: Do you really need audio to fine-tune your audio llm? *CoRR*, abs/2505.09439, 2025.
- S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.
- Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, and Shiguo Lian. DAST: difficulty-adaptive slow-thinking for large reasoning models. *CoRR*, abs/2503.04472, 2025.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*, 2025.
- Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. *CoRR*, abs/2503.02318, 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.
- Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. Text-to-audio grounding: Building correspondence between captions and sound events. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren,

Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025.

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 *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. ACM, 2022.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*. Association for Computational Linguistics, 2024.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightweight infrastructure for fine-tuning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*. AAAI Press, 2025.

## A APPENDIX

### A.1 TASK DEFINITION

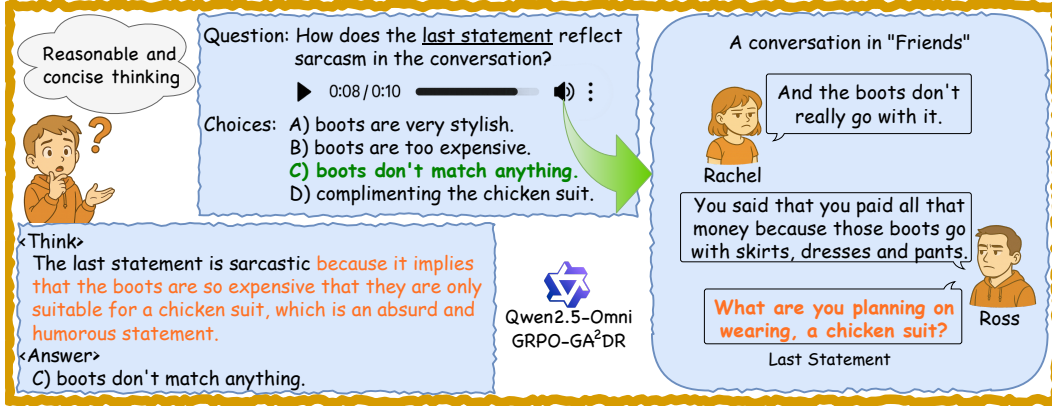


Figure 4: An audio QA example from “Friends”. The top-left shows the question and options (green indicates the correct one), the right side presents the audio dialogue, and the bottom-left shows the output of our proposed method on Qwen2.5-Omni-7B.

### A.2 SETUP

**Models.** For the base models, we primarily adopt Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B, which are among the most widely used open-source LALMs. These two models are also frequently used in prior work, facilitating fair and direct comparisons. For the training framework, we leverage the ms-swift (Zhao et al., 2025) and perform training on three A100-40G GPUs, where two GPUs are used for model training and one GPU is reserved for vLLM-based inference (Kwon et al., 2023). More training details and hyper-parameters can be found in the Appendix A.8.

**Datasets.** For the training datasets, we use two in total: FS and AVQA (Yang et al., 2022). The latter is a subset of the former, while FS is constructed by augmenting AVQA with four additional datasets. For AVQA, we only keep the audio-text pairs and replace the word “video” in the questions with “audio.” For evaluation, we mainly test on MMAU-test-mini (Sakshi et al., 2025), with AirBench Foundation-Sound (Yang et al., 2024) as a secondary reference.

**GRPO.** GRPO has been extensively applied in both LLMs and MLLMs, achieving notable progress, and our implementation largely follows prior studies (DeepSeek-AI et al., 2025; Li et al., 2025). Compared with other RL methods, the key feature of GRPO is that it evaluates the policy model’s advantage using the average reward of in-group sampled outputs. Given an input question, a set of sampled responses for that question, and their rewards from the reward function, the advantage is calculated as follows:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (6)$$

Here,  $A_i$  denotes the advantage used to optimize the policy model,  $\{r_1, r_2, \dots, r_G\}$  represents the set of reward values corresponding to each sampled output within the group  $\{o_1, o_2, \dots, o_G\}$ , and  $G$  indicates the number of samples in the group.

After this, GRPO uses the computed advantage to optimize the policy model by maximizing the following objective function:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)] \quad (7)$$

$$\frac{1}{G} \sum_{i=1}^G \left( \min \left( \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left( \frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right) A_i \right) - \beta \mathcal{D}_{KL}(\pi_{\theta} \parallel \pi_{ref}) \right).$$

$$\mathbb{D}_{KL}(\pi_{\theta} \parallel \pi_{ref}) = \frac{\pi_{ref}(o_i | q)}{\pi_{\theta}(o_i | q)} - \log \frac{\pi_{ref}(o_i | q)}{\pi_{\theta}(o_i | q)} - 1. \quad (8)$$

Here,  $\varepsilon$  and  $\beta$  are hyper-parameters.

**Cold-start GRPO.** A common way to learn the reasoning structure of advanced models and further optimize performance is to distill their outputs and then apply SFT to quickly teach the model such paradigms. After this, GRPO is performed, and this approach is referred to as Cold-Start. To explore the effectiveness of this method, we also conduct experiments with SFT Cold-Start followed by GRPO. For the Cold-Start dataset, we first sample a subset from the FS training set, distill it with Gemini2.5Pro, and then use Qwen3-235B-A22B<sup>3</sup> to retain only samples that are both correct and consistent in reasoning and answers. From these, we select 200 per task to form a dataset of 1,000 samples. This dataset is then used to perform Cold-Start SFT on Qwen2-Audio-7B-Instruct, and the model obtained after 2 epochs of SFT serves as the starting checkpoint for GRPO.

**Implicit and Explicit Prompt.** To examine whether the effectiveness of CoT comes from explicit outputs or simply triggering implicit reasoning, we design two prompts, shown in Figure 5. For each sample, the model generates its answer within `<answer>` `</answer>`. In Prompt1, CoT is not required, while in Prompt2, it is generated within `<think>` `</think>`. A Format-Reward enforces this structure in Prompt2, whereas in Prompt1 it only regulates the answer format.

**Prompt Templates**

Prompt1: Output the most suitable answer (from A, B, C, and D with its corresponding answer) in `<answer>` `</answer>` tags.

Prompt2: Output your thinking process in `<think>` `</think>` and put the most suitable answer (from A, B, C, and D with its corresponding answer) in `<answer>` `</answer>` tags.

Figure 5: Different prompt templates for GRPO, where Prompt1 is the implicit prompt and Prompt2 is the explicit prompt.

<sup>3</sup><https://huggingface.co/Qwen/Qwen3-235B-A22B>

**Truncation Reward.** In addition to our designed rewards, we also consider the most straightforward way to control reasoning length as one of the baseline methods. This strategy sets a length threshold  $L_T$ , where rollout samples that are correct and have length less than or equal to  $L_T$  receive a positive reward, while those exceeding it receive a penalty, even if the answer is correct. The calculation is as follows:

$$r_i = \begin{cases} 1, & \text{if } L_{o_i} \leq L_T \text{ and } o_i \text{ is correct,} \\ \sigma, & \text{if } L_{o_i} > L_T. \end{cases} \quad (9)$$

Here,  $L_{o_i}$  denotes the output length of the CoT for the corresponding sample, and  $\sigma$  is a penalty hyper-parameter.

### A.3 MODEL-PERSPECTIVE DIFFICULTY ON MMAU

In Sakshi et al. (2025), question difficulty was annotated manually, where multiple experts assigned difficulty scores to each question. These labels are of very high quality, but the cost of such annotation is prohibitively high, and the labels are fixed, making it difficult to align them with the model’s evolving state across different training steps.

Therefore, in this subsection, we introduce model-perspective difficulty labels. Since each model has different capabilities, we aim to reflect an average level across models. To this end, we adopt four different models: Qwen2-Audio-7B-Instruct, Qwen2.5-Omni-7B, Kimi-Audio-7B-Instruct<sup>4</sup>, and Gemini2.5-Pro-0506. These models perform inference on MMAU-test-mini under the same random seed, and difficulty labels from the model’s perspective are assigned based on the number of models answering each question correctly. The detailed distribution of model-perspective difficulty labels is shown in Table 9.

Table 9: Data distribution of difficulty from human (Orig) and model perspectives, including counts of changed (Chg) and unchanged (Un-Chg) samples, and transitions across difficulty categories.

Orig Diff.	Total-Num		Num		New Diff.	Chg Num
	Orig	New	Un-Chg	Chg		
Easy	258	527	97	161	Medium	68
					Hard	93
Medium	510	214	91	419	Easy	338
					Hard	81
Hard	232	259	85	147	Easy	92
					Medium	55

### A.4 IMPACT OF THRESHOLD RATIO ON OUR PROPOSED REWARD

Considering that when the negative exponential function approaches 1 its exponent tends toward 0, this can represent a form of implicit reasoning, though it may reduce readability. To further analyze this issue, we introduce a parameter  $l_{min}$ , which serves as the minimum threshold ratio. When the relative length of a sample is smaller than this value, it is directly set to 1; when it is larger, its length is further normalized. The  $\zeta$  then applies this secondary normalization after the threshold.

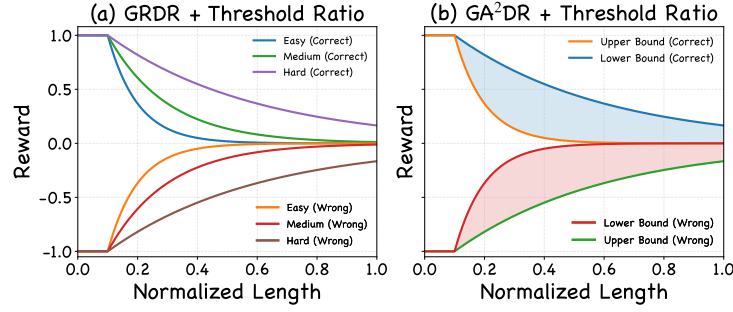
$$r_i = \text{sign}(o_i) \cdot e^{-k(\gamma)\zeta(l_{o_i}; l_{min})} \quad (10)$$

$$\zeta(l_{o_i}; l_{min}) = \max(0, \frac{l_{o_i} - l_{min}}{1 - l_{min}}) \quad (11)$$

Table 10, Table 11, and Table 12 report the effects of adding a threshold ratio to our two proposed rewards on three benchmarks. Models without special notation correspond to the case of  $l_{min} = 0$ , which means that no threshold ratio is applied. For the cases with  $l_{min} = 0.1$ , this value is determined based on the average reasoning length of direct GRPO models.

<sup>4</sup><https://huggingface.co/moonshotai/Kimi-Audio-7B-Instruct>



Figure 6: Curves of GRDR and GA<sup>2</sup>DR with normalized length and the threshold ratio  $l_{min}$ .Table 10: The performance of our proposed GRDR, GA<sup>2</sup>DR, and their variants with added length threshold ratios on MMAU-Test-Mini. Here, unmarked models correspond to  $l_{min} = 0$ , and  $\dagger$  denotes the model-perspective difficulty annotations.

Models	MMAU-Test-Mini									
	Sound	Music	Speech	Easy	Easy <sup>†</sup>	Medium	Medium <sup>†</sup>	Hard	Hard <sup>†</sup>	Avg
Based On Qwen2.5-Omni-7B										
GRDR	71.47	<b>72.45</b>	66.66	<b>60.07</b>	93.16	80.00	58.87	59.91	<b>32.81</b>	<b>70.20</b>
+ $l_{min} = 0.1$	<b>72.07</b>	69.76	63.36	55.81	<b>93.92</b>	79.60	55.14	57.75	27.41	68.40
GA <sup>2</sup> DR	71.47	71.85	<b>66.66</b>	57.75	92.78	<b>80.58</b>	<b>59.81</b>	<b>60.34</b>	32.04	70.00
+ $l_{min} = 0.1$	71.77	68.86	64.26	55.42	93.73	79.02	57.00	59.05	25.86	68.20

Table 11: The performance of our proposed rewards, and their variants with added length threshold ratios on MMAU-v0515. Here, unmarked models correspond to  $l_{min} = 0$ .

Models	MMAU-Test-Mini-v0515			
	Sound	Music	Speech	Avg
Based On Qwen2.5-Omni-7B				
GRDR	<b>83.48</b>	70.35	75.97	76.60
+ $l_{min} = 0.1$	83.18	70.95	76.27	76.80
GA <sup>2</sup> DR	83.18	71.55	75.67	76.80
+ $l_{min} = 0.1$	81.08	<b>72.15</b>	<b>77.77</b>	<b>77.00</b>

Table 12: The performance of our proposed rewards, and their variants with added length threshold ratios on MMAR. Here, unmarked models correspond to  $l_{min} = 0$ .

Models	MMAR			
	Sound	Music	Speech	Avg
Based On Qwen2.5-Omni-7B				
GRDR	61.21	51.94	65.30	61.20
+ $l_{min} = 0.1$	63.63	52.91	<b>65.98</b>	63.00
GA <sup>2</sup> DR	64.84	<b>54.85</b>	65.30	62.90
+ $l_{min} = 0.1$	<b>64.84</b>	53.39	63.60	<b>63.00</b>

From an overall performance perspective, models with a length threshold ratio achieve higher average scores on MMAU-Test-Mini-v0515 and MMAR, but show the opposite trend on MMAU-Test-Mini. Considering that the quality of the Q&As and audio files in MMAU-Test-Mini is relatively poor, the performance on the latter two benchmarks is therefore more convincing. The fact that introducing a length threshold leads to better results on most benchmarks further indicates that our length-based reward curve tends to infinitely optimize toward a completion length of zero, which may cause the CoT content of certain samples to become less effective, ultimately resulting in a decrease in overall performance.

From the perspective of different difficulty levels, this part mainly focuses on the results from the two versions of MMAU-Test-Mini. We observe that the model with the added threshold performs better on easy and hard questions, while the model without the threshold achieves better results on medium questions. This outcome is closely related to our threshold setting: the threshold value was determined based on the average reasoning length of the direct GRPO model on MMAU-Test-Mini. Since more than half of the questions in the dataset are labeled as medium difficulty, the threshold essentially aligns with this group. As a result, medium questions—which could have

benefited from further length optimization—received less optimization, leading to relatively lower performance. In contrast, for easy and hard questions, introducing the threshold brings two benefits: for easy questions, it prevents excessively short reasoning that may reduce CoT effectiveness; for hard questions, it increases the upper bound of reasoning length, enabling deeper exploration of problem-solving strategies. Consequently, the model achieves better performance on both easy and hard questions.

In addition, when focusing on GRDR and GA<sup>2</sup>DR, we can observe that the latter achieves better performance on two benchmarks. This is because GA<sup>2</sup>DR defines difficulty levels at the batch level, which aligns with the batch-wise optimization process during model backpropagation, making this approach more effective. Moreover, its attention mechanism jointly considers both the textual question and the characteristics of the audio itself.

Furthermore, previous studies (Peng et al., 2025) have shown that models trained with reinforcement learning tend to exhibit improvements in pass@1 but declines in pass@k, largely due to the instability of rollout samples. Our GA<sup>2</sup>DR method effectively mitigates this issue, since all rollout samples within a group share the same question and audio input, thereby avoiding inconsistencies caused by rollout variance.

In summary, combining GA<sup>2</sup>DR with a length threshold provides the most balanced and effective optimization strategy. The specific threshold value, however, should be determined based on the characteristics of each individual task.

#### A.5 DETAILED LENGTH STATISTICS OF DIFFERENT MODELS ON MMAU

In this section, we provide the raw reasoning-length data for all models used in our main experiments on MMAU-Test-Mini, without any post-processing. All lengths are measured using the Qwen2-Audio tokenizer. Note that Qwen2-Audio and Qwen2.5-Omni share the same tokenizer, so using the former does not introduce any error when evaluating the latter. For Gemini2.5, however, we cannot obtain its tokenizer because it is a closed-source model. Although using the Qwen2-Audio tokenizer may introduce some deviation, the reasoning lengths of Gemini2.5 are far greater than those of all our methods, so this small discrepancy does not affect the clearly observable relative differences.

Table 13: The output length statistics of models on MMAU-test-mini with different base models and reward settings, reported under both the human-perspective difficulty annotations and the assigned model-perspective difficulty annotations. Here,  $\dagger$  denotes the model-perspective difficulty annotations.

Models	MMAU-Test-Mini					
	Easy	Easy <sup>†</sup>	Medium	Medium <sup>†</sup>	Hard	Hard <sup>†</sup>
Baseline Models						
Qwen2-Audio-7B-Instruct	59.79	65.48	67.66	65.26	70.18	68.50
Qwen2.5-Omni-7B	151.80	89.76	80.19	121.90	121.57	134.67
Kimi-Audio-7B-Instruct	94.99	21.79	29.69	59.03	25.01	82.38
Gemini2.5-Pro-0506	759.60	525.85	542.17	663.58	628.32	768.81
Based On Qwen2-Audio-7B-Instruct						
GRPO	41.76	42.26	42.55	43.29	43.89	42.97
+ TR	42.11	43.55	43.77	44.42	46.32	44.31
+ GRDR	27.76	30.41	31.59	31.49	32.26	30.88
Based On Qwen2.5-Omni-7B						
GRPO	185.38	108.08	99.74	153.90	146.18	164.92
+ TR	183.65	110.20	103.10	153.49	145.76	165.46
+ GRDR	54.49	51.69	41.70	56.05	68.32	56.73
+ $l_{min} = 0.1$	124.62	89.10	85.66	110.00	107.54	116.96
+ GA <sup>2</sup> DR	54.58	52.84	48.57	53.92	66.48	57.58
+ $l_{min} = 0.1$	115.38	88.80	85.34	105.89	105.91	109.68

## A.6 REASONING LENGTH VS. ACCURACY

In this section, we analyze the relationship between output length and accuracy for the GRDR and GA<sup>2</sup>DR methods on MMAU-Test-Mini. From Figure 7 to Figure 10, samples are grouped into length intervals of similar size, and correctly answered samples within each interval are further divided by difficulty level, indicated by increasingly darker shades of blue.

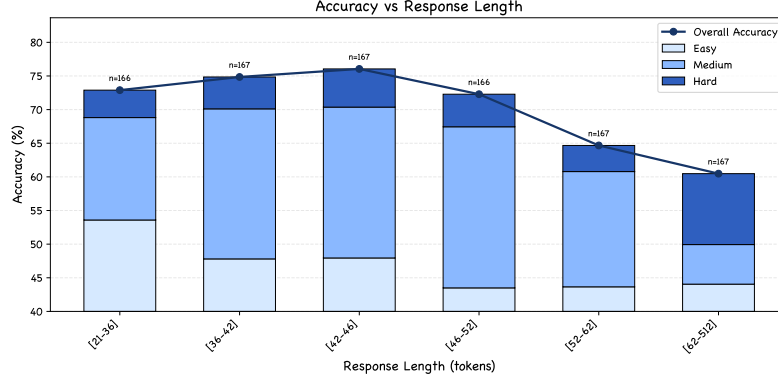


Figure 7: The trend of length and accuracy for the GRDR with human-perspective difficulty.

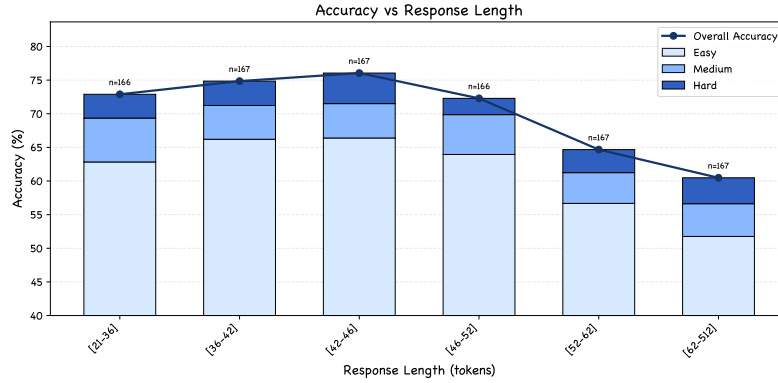


Figure 8: The trend of length and accuracy for the GRDR with model-perspective difficulty.

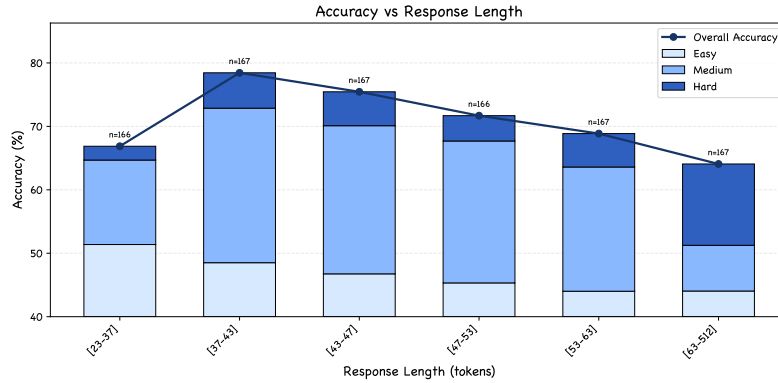


Figure 9: The trend of length and accuracy for the GA<sup>2</sup>DR with human-perspective difficulty.

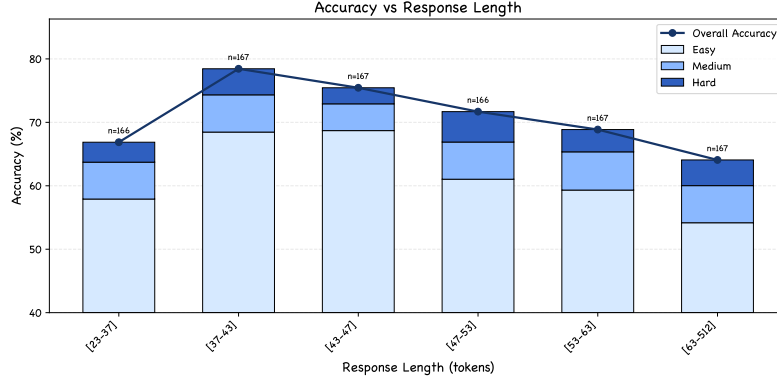


Figure 10: The trend of length and accuracy for the GA<sup>2</sup>DR with model-perspective difficulty.

### A.7 TRAINING CURVES

In this section, we present the training curves of GRDR and GA<sup>2</sup>DR on Qwen2.5-Omni, including grad norm, KL divergence, and reward trends in Figure 11 and Figure 12. All curves are directly extracted from the TensorBoard logs recorded during training, with a smoothing factor of 0.8 to clearly reveal their variations and overall trajectories.

Overall, both methods converge stably without exhibiting extreme values in gradients or KL divergence, and no numerical spikes are observed. This indicates that our training procedures are stable and that the proposed methods are well-behaved in practice.

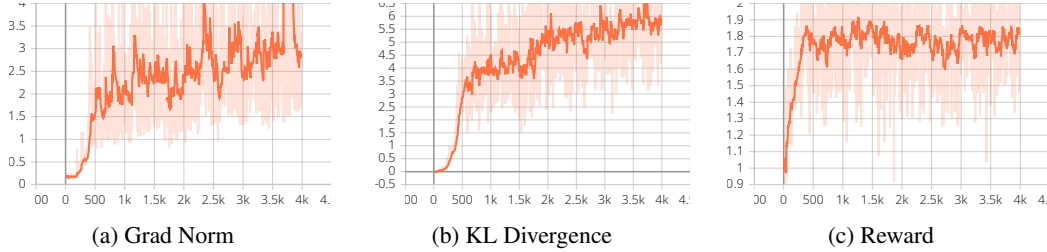


Figure 11: Training curves of gradient norm, KL divergence, and reward on GRDR.

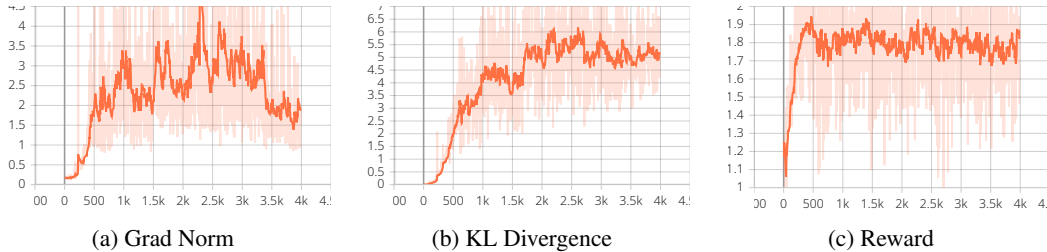


Figure 12: Training curves of gradient norm, KL divergence, and reward on GA<sup>2</sup>DR.

### A.8 HYPER-PARAMETERS

In this part, we provide detailed explanations of the hyperparameter settings in Table 14, including the two proposed rewards and the basic TR hyperparameter settings.

Table 14: Hyper-parameters for SFT and GRPO, including different settings under LoRA and Full.

LoRA Fine-tuning	
LoRA_rank	32
LoRA_alpha	32
Torch_dtype	bfloat16
Max_length	1024
Num_train_epochs	1
Per_device_train_batch_size	8
Per_device_eval_batch_size	8
Gradient_accumulation_steps	16
Learning_rate	1.5e-6
Num_generations	8
Temperature	1.0
Warmup_ratio	0.03
Beta	0.04
Epsilon	0.2
Deepspeed	zero2
Rule-base Reward	
Truncation Reward $L_T$	120 / 400
Truncation Reward $\sigma$	-0.5
GRDR and GA <sup>2</sup> DR $l_{min}$	0.1
GRDR and GA <sup>2</sup> DR $k_{hard}$	2
GRDR and GA <sup>2</sup> DR $k_{easy}$	10

#### A.9 CASE STUDY: QUALITATIVE ANALYSIS OF OUTPUT PARADIGMS ACROSS MODELS

In this section, we analyze output paradigms across models using 20 MMAU-test-mini questions on Dissonant Emotion Interpretation, focusing on sarcasm cause detection. The corresponding audios often involve multiple speakers and complex environments, making the task challenging and well-suited for qualitative analysis of different model outputs. In this section, in addition to the main models discussed above, we also include a Cold-Start GRPO model based on Qwen2-Audio-7B-Instruct, with detailed settings provided in the Appendix A.2.

The detailed results are shown in Table 15, which include ACC and reasoning length on the subset mentioned above. In the second part, “Based on Qwen2-Audio-7B-Instruct,” we compare our three GRPO models with Audio-Reasoner. GRPO and GRPO with GRDR match Audio-Reasoner in performance while producing much shorter outputs, whereas the GRPO model after Cold-Start produces reasoning lengths close to Audio-Reasoner but performs worst. We believe this is mainly because the base model is weaker and can only imitate the surface structure of advanced model paradigms, showing that learning paradigms through Cold-Start is not effective in all cases and can even yield negative effects. In the “Based on Qwen2.5-Omni-7B” part of the table, performance differences are minor, mainly because the Qwen2.5-Omni base model already has some reasoning ability and shows good performance on complex questions. However, with our two proposed rewards, the models maintain their original performance while significantly shortening reasoning length, greatly improving reasoning efficiency.

In addition to analyzing performance and reasoning length above, we also compared the outputs of different models on these 20 questions in detail. From the results, we found that models with better performance usually follow a complete reasoning process, which includes first grounding the audio and providing a corresponding caption—equivalent to identifying the known conditions in the problem—then reasoning based on these conditions, analyzing each option step by step, and finally giving the answer. This approach helps LALMs make better use of known information together with their broad pretrained knowledge to perform reasonable reasoning while also improving readability. However, weaker models, even when using this structure, often make mistakes in the first step of extracting known information. These errors propagate through the reasoning process, leading to wrong answers and lower performance. For this part, we also provide an example of different models’ outputs in Appendix A.10 as reference.

Table 15: Performance and reasoning length of different models on 20 sarcasm cause detection tasks.

Model	ACC	Avg-Length
Advanced Proprietary Model		
Gemini2.5-Pro-0506	95	931.8
Based On Qwen2-Audio-7B-Instruct		
Audio-Reasoner	75	547.1
GRPO	75	50.7
+ GRDR	70	38.5
+ Cold-Start SFT	55	541.4
Based On Qwen2.5-Omni-7B		
GRPO	100	109.5
+ GRDR	95	56.2
+ GRDR ( $l_{min} = 0.1$ )	90	94.1
+ GA <sup>2</sup> DR	100	56.9
+ GA <sup>2</sup> DR ( $l_{min} = 0.1$ )	95	94.2

Overall, we believe that a good reasoning paradigm should consist of three parts: extracting known conditions, reasoning and analysis, and giving the final answer. Among these steps, the earlier ones have a stronger impact on later results, and once errors occur, they are difficult to correct in the following steps. Therefore, we suggest that for weaker models, it is better to first fine-tune them on basic tasks related to extracting known conditions (e.g., grounding and caption) before learning reasoning paradigms. For stronger models that still produce reasoning with poor readability, Cold-Start can be used to learn the paradigm first, followed by RL training. For models that already generate reasoning with good structure, direct RL with our two proposed rewards can further improve both performance and efficiency.

#### A.10 EXAMPLE ANALYSIS

In this section, we present output examples from six different models on a sarcasm cause detection task from Figure 13 to Figure 15. This question is labeled as Medium in the original MMAU-test-mini but is relabeled as Hard under our new scheme. The models include an advanced proprietary model, Gemini2.5-Pro, the prior work Audio-Reasoner, and the four main models used in our experiments. In the figure, the green text corresponds to the correct option for this example, the blue parts in each model’s reasoning indicate logical and necessary steps or key points for solving the question, and the red parts indicate repetition or redundancy, representing the parts that need optimization.

**Example Analysis**

Question: How does the last statement reflect sarcasm in the conversation?  
 Choices: A) boots are very stylish. B) boots are too expensive.  
 C) boots don't match anything. D) complimenting the chicken suit.

 0:08 / 0:10 

**Gemini2.5-Pro**

<Think>  
 The conversation revolves around ... The woman ... she states that "... (referring to an outfit she's considering). The man responds sarcastically, "..." Let's analyze ... 1. The man is highlighting ... She claimed ... 2. The "..." is an absurd and ridiculous outfit. Suggesting ... 3. The sarcasm is ... Now let's evaluate the choices: A) The ... doesn't directly comment on ... B) The expense is mentioned ... but ... C) The man's "..." This interpretation suggests ... So ... D) This is clearly incorrect ... Let's consider ... more deeply. The man is saying, "..." The sarcasm points to ... If the boots ... suggests ... This could imply that ... Thus, the implication ... The chicken suit serves as ... By suggesting ... If ... then ... Alternatively, if ... then ... The sarcasm plays on this dilemma. Option C ... The answer is C.

<Answer>  
 C) boots don't match anything.

**Audio Reasoner**

<Think>  
 <PLANNING>: The question asks ... I need to analyze ... Then ... I will focus on ...  
 <CAPTION>: The audio ... The first speaker mentions "..." The second speaker responds "..." The first speaker's tone ... and the second speaker's tone ...  
 <REASONING>: The question asks ... Let's ... : (a): ... is unlikely. (b): ... is less likely. (c): ... This option aligns well ... (d): ... is incorrect. Therefore ... (c).  
 <SUMMARY>: The second speaker's ... "..." implies that ... This aligns with option (c).

<Response>  
 The last statement ... implying that ... The answer is C.

**Qwen2-Audio GRPO**

<Think>  
 The speaker is sarcastic because they imply that the boots are only meant to be worn with certain outfits, which contrasts with the previous statement implying they were versatile.

<Answer>  
 C) boots don't match anything.

**Qwen2-Audio Cold-Start GRPO**

<Think>  
 The audio starts "..." This implies that ... Then ... This reinforces the idea that ... Finally ... This implies that ... Let's evaluate the choices: A) ... this is not ... B) This is also not necessarily ... C) This aligns well with ... D) This is clearly ... The last statement is "..." This is ... implying that ... Therefore ... The final answer is C.

<Answer>  
 C) boots don't match anything.

**Qwen2.5-Omni GRPO**

<Think>  
 The last statement, "a chicken suit?" is a rhetorical question that implies the boots are not suitable for any outfit, including a chicken suit. This is a sarcastic remark, as it is unlikely that the boots would be appropriate for a chicken suit. The other options do not reflect the sarcasm in the statement. Therefore, the most suitable answer is C.

<Answer>  
 C) boots don't match anything.

**Qwen2.5-Omni GRPO-GA<sup>2</sup>DR**

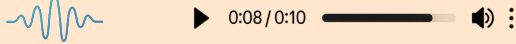
<Think>  
 The last statement is sarcastic because it implies that the boots are so expensive that they are only suitable for a chicken suit, which is an absurd and humorous statement.

<Answer>  
 C) boots don't match anything.

Figure 13: Output examples from six different models on a sarcasm cause detection question. The green part in the question indicates the correct option. The blue parts in the model outputs represent core content, while the red parts indicate redundancy.

**Example Analysis**

Question: Why can the last line be interpreted as sarcastic?  
 Choices: A) They love assembling furniture together. B) They don't know Joey and Chandler.  
 C) They are not coming over. D) Not genuinely excited about assembling.



**Gemini2.5-Pro**

<Think>  
 The audio clip presents a conversation. Speaker 1: ... Speaker 2: ... The question asks why the last line ... can ... Sarcastm is ... It often involves ... with a tone ... Let's analyze Speaker 2's line and tone ... Now let's evaluate the given choices: A) ... So, A is unlikely. B) ... C) ... So, C is not the most direct reason for the sarcasm. D) ... This matches option D. Let's consider why the other options are less likely ... Therefore, the most suitable reason ... Re-listening to the audio ... Therefore, the sarcastic interpretation ... Final Answer seems to be D. The final answer is D.

<Answer>  
 D) Not genuinely excited about assembling.

**Audio Reasoner**

<Think>  
 <PLANNING>: The question asks ... I need to analyze ... Then ... I will focus on ...  
 <CAPTION>: The audio clip features a conversation where ... The speaker mentions ... The speaker then says ... The speaker's tone is ... and the phrase .. suggests ...  
 <REASONING>: Let's ... (a) ... (b) ... (c) ... (d) ... Therefore, the most ... is not coming over.  
 <SUMMARY>: The speaker's tone ... The most ... Therefore, the answer is C.

<Response>  
 The speaker's tone ... The most ... The answer is C.

**Qwen2-Audio GRPO**

<Think>  
 The speaker might be sarcastic because they don't feel like being alone tonight but is forced to invite others over, implying their true feelings.

<Answer>  
 C) They are not coming over.

**Qwen2-Audio Cold-Start GRPO**

<Think>  
 The audio contains the following lines: "...". This statement implies that ... Now let's consider the choices: A) ... B) ... C) ... D) ... Considering the options, "...". seems like the most plausible reason for sarcasm. If the speaker were genuinely excited ... they would likely ... Instead, they mention ... The final answer is B.

<Answer>  
 B) They don't know Joey and Chandler.

**Qwen2.5-Omni GRPO**

<Think>  
 The last line ... could be interpreted as sarcastic because it implies that the speaker is not genuinely excited about the offer to help assemble furniture. The use of ... suggests a dismissive or sarcastic tone, indicating that the speaker is not really thankful for the offer. Therefore, the most suitable answer is D).

<Answer>  
 D) Not genuinely excited about assembling.

**Qwen2.5-Omni GRPO-GA<sup>2</sup>DR**

<Think>  
 The last line is sarcastic because it implies that the speaker is not genuinely excited about assembling furniture. The speaker is thanking someone sarcastically, which suggests that they are not actually excited about the situation.

<Answer>  
 D) Not genuinely excited about assembling.

Figure 14: Output examples from six different models on a sarcasm cause detection question. The green part in the question indicates the correct option. The blue parts in the model outputs represent core content, while the red parts indicate redundancy.



**Example Analysis**

Question: What makes the last comment sarcastic in relation to the dialogue?  
 Choices: A) She doesn't like birthdays. B) She won't remember anyway.  
 C) She's too old for parties. D) We can reschedule her birthday.

 0:08 / 0:10 

**Gemini2.5-Pro**

<Think>  
 The dialogue involves a disagreement about ... One person insists on ... The last speaker makes a sarcastic comment: ... Let's analyze what makes this comment sarcastic in relation to the dialogue: The speaker ... Instead ... If Emma were ... and had no idea ... several things would be true:1. She wouldn't ... 2. She wouldn't ... 3. Her memory ... The sarcasm implies that ... By wishing ... The sarcastic implication is that ... Now let's evaluate the choices:A) ... B) ... C) ... D) ... It's a specific jab ... The sarcasm hinges on the characteristics ... A key characteristic related ... If Emma were ... The speaker is sarcastically ... This is used to imply ... Therefore, the fact ... The sarcastic comment ... The most suitable answer is B because ... The final answer is B

<Answer>  
 B) She won't remember anyway.

**Audio Reasoner**

<Think>  
 <PLANNING>: The question asks what ... I need to analyze ... Then, I will evaluate ...  
 <CAPTION>: The audio ... The person ... The speaker then says ... This is followed by laughter.  
 <REASONING>: The core of the sarcasm lies in ... The person ... The comment ... is ... implying ... The laughter at the end confirms ... Now, let's evaluate the options:A) ... B) ... C) ... D) ... Therefore, option C is ...  
 <SUMMARY>: The sarcasm ... arises from ... This is ... Therefore, the answer is C.

<Response>  
 The last comment ... implies that ... The answer is C.

**Qwen2-Audio GRPO**

<Think>  
 The speaker seems to express disbelief or resignation regarding the importance of birthdays, especially when compared to the ongoing party plans.

<Answer>  
 B) She won't remember anyway.

**Qwen2-Audio Cold-Start GRPO**

<Think>  
 The audio contains a conversation where one person ... The other person ... This statement implies ... Let's evaluate the choices: A) ... B) ... C) ... D) ... Given the context, the statement ... is the most likely source ... It suggests that the person ... The final answer is B.

<Answer>  
 B) She won't remember anyway.

**Qwen2.5-Omni GRPO**

<Think>  
 The last comment is sarcastic because it implies that the person being addressed is too old to be surprised by the concept of a birthday party. The speaker is suggesting that the person is old enough to know what a birthday is, which is a form of sarcasm. The other options do not fit the context of the dialogue as well as B does.

<Answer>  
 B) She won't remember anyway.

**Qwen2.5-Omni GRPO-GA<sup>2</sup>DR**

<Think>  
 The speaker is expressing frustration about the situation. The sarcastic comment implies that the speaker wishes the child was too young to remember birthdays, which is not true.

<Answer>  
 B) She won't remember anyway.

Figure 15: Output examples from six different models on a sarcasm cause detection question. The green part in the question indicates the correct option. The blue parts in the model outputs represent core content, while the red parts indicate redundancy.