Think Smart, Not Hard: Difficulty Adaptive Reasoning for Large Audio Language Models

Anonymous ACL submission

Abstract

Large Audio Language Models (LALMs) employing the Chain-of-Thought paradigm have demonstrated remarkable reasoning capabilities. Though different problems naturally require varying depths of reasoning, existing methods often determine whether to perform reasoning, lacking fine-grained mechanisms to adapt reasoning length to problem complexity. As a result, LALMs often adopt a one-size-fits-all reasoning strategy, leading to redundant overthinking for simple tasks and insufficient reasoning for complex ones. In this paper, we conduct an in-depth analysis of LALM reasoning behavior and argue that effective and efficient reasoning should be adaptively aligned with task difficulty. To this end, we propose a difficulty-adaptive reasoning method for LALMs. Specifically, we introduce a reward function that dynamically links reasoning length to the model's perceived problem difficulty, encouraging shorter reasoning for easy tasks and longer reasoning for more complex ones. Extensive experiments on three datasets demonstrate that our method consistently improves performance while reducing average reasoning length by at least 50%, achieving higher efficiency without sacrificing accuracy.

1 Introduction

Large Audio Language Models (LALMs) (Tang et al., 2024; Chu et al., 2024), powered by the chain-of-thought (CoT) (Xie et al., 2025; Li et al., 2025; Wu et al., 2025), have advanced rapidly in recent years, demonstrating strong capabilities in audio understanding and reasoning. However, such improvements often rely on increasingly long and exhaustive reasoning processes. This raises an important question: how can LALMs reason more effectively while avoiding unnecessary reasoning?

Prior work by Xie et al. (2025) enables reasoning in LALMs through supervised fine-tuning (SFT) on large-scale datasets with CoT annotations, demonstrating the effectiveness of explicit reasoning for complex audio question answering. Subsequently, Li et al. (2025) introduce group relative policy optimization (GRPO) into LALMs, achieving better performance than SFT without requiring CoT-annotated datasets. However, these studies lack a systematic analysis of how different strategies perform under varying levels of question difficulty. Motivated by this gap, in this work, we conduct a comprehensive analysis and observe that GRPO performs better on harder questions, while being slightly inferior on easier ones. Moreover, when comparing GRPO with explicit and implicit prompts, explicit reasoning shows clear advantages on difficult questions and achieves comparable performance to the implicit setting on easier ones. Taken together, these findings suggest that efficient reasoning in LALMs requires adapting reasoning length to problem difficulty, rather than applying a uniform reasoning strategy.

Regarding reasoning length and efficiency, Qu et al. (2025) identify issues such as redundancy and overthinking in reasoning. Existing RL-based studies on LLMs (Arora and Zanette, 2025; Aggarwal and Welleck, 2025) adopt length-penalty rewards, but rely on fixed thresholds without accounting for differences in question difficulty. For LALMs, Wu et al. (2025) propose a "when to think" mechanism to decide whether reasoning is needed; however, it does not regulate reasoning length once reasoning is required. These limitations indicate that an effective LALM should adjust reasoning length according to task complexity, using shorter, concise reasoning for simple questions and longer, deeper reasoning for harder ones. Motivated by this insight, we propose a length based reward that avoids fixed thresholds and introduce two difficulty-adaptive reasoning methods for LALMs.

In this work, we evaluate LALMs on the audio question answering task. As illustrated in Figure 1, given an audio clip, a corresponding question, and

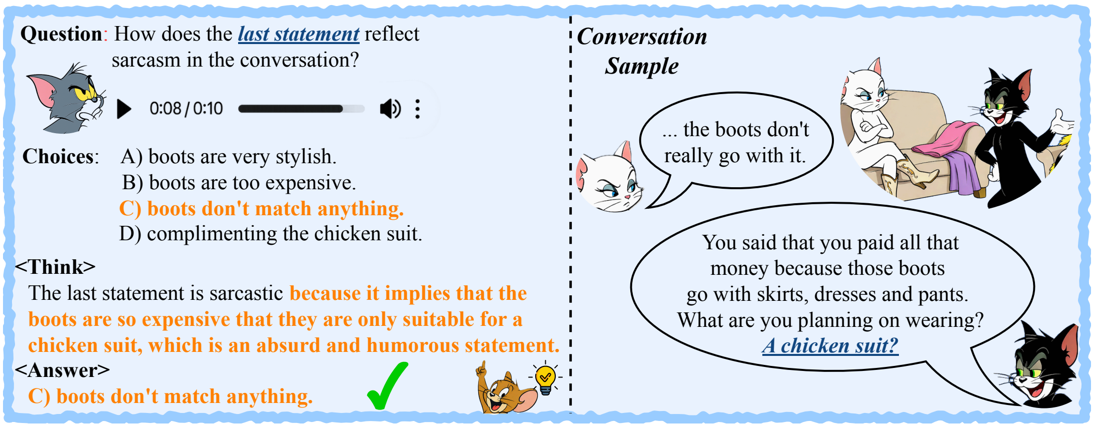


Figure 1: An audio Q&A example from "Tom and Jerry". The left shows the question, candidate choices, CoTs and the final answer, the right side presents the dialogue.

four candidate answers as the input, the model is required to select the correct answer through reasoning. Experiments on MMAU (Sakshi et al., 2025), MMAU-v0515, and MMAR (Ma et al., 2025) show that the proposed rewards get strong performance with much shorter reasoning length than direct GRPO models, demonstrate their effectiveness.

Overall, the main contributions are as follows:

- We conduct in-depth analyses of LALMs and show that effective LALMs should reason smartly, with reasoning length adjusted to match problem difficulty.
- We propose two difficulty-adaptive, length-based rewards that reduce overall reasoning length while enabling shorter reasoning for simple questions and longer reasoning for more difficult ones.
- We perform extensive experiments on three datasets, demonstrating that both proposed difficulty-adaptive rewards consistently improve performance while reducing reasoning length by about 50%, thereby achieving simultaneous gains in effectiveness and efficiency.

We have also provided an anonymous repository at `https://anonymous.4open.science/r/ARR-January-Anonymous-Repo-2958` to further support the reproduction of this work.

2 Deep Analysis of Different Methods for LALMs

Recent studies (Xie et al., 2025; Li et al., 2025) enhance LALMs' ability through reasoning but lack detailed analysis of the sources of these gains.

Task	Dataset-Source	Num
Audio Grounding	AudioGrounding	1,805
Sound Classification	VocalSound	15,531
	TUT2017	3,744
Sound Question Answering	Clotho-AQA	6,615
	AVQA	36,036

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

Therefore, we analyze this from two perspectives. We first compare SFT and GRPO under different conditions and then examine whether improvements arise from explicit or implicit prompting.

Data. For training, we use FS and AVQA (Yang et al., 2022). FS extends AVQA with four additional datasets covering three tasks, including AudioGrounding (Xu et al., 2021), VocalSound (Gong et al., 2022), TUT2017 (Mesaros et al., 2017), and Clotho-AQA (Lipping et al., 2022). Thus, FS is approximately twice the size of AVQA, with detailed statistics shown in Table 1.

Setup. We use Qwen2-Audio-7B-Instruct[1] and Qwen2.5-Omni-7B[2] as base models. Experimental settings are provided in Appendix A.1. Table 2 reports all evaluations using accuracy, including averages for methods with identical settings.

[1]https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct
[2]https://huggingface.co/Qwen/Qwen2.5-Omni-7B

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, Prompt 2	81.10	67.57	64.67	62.76	55.04	73.73	56.90	65.00
On AVQA, Full, Prompt 2	70.35	67.87	66.77	60.96	52.33	75.10	57.76	65.20
On FS, LoRA, Prompt 2	69.60	69.37	60.48	55.26	49.61	71.76	53.02	61.70
On AVQA, LoRA, Prompt 2	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, Prompt 1	69.87	67.27	61.98	56.76	50.00	71.57	54.31	62.00
On AVQA, LoRA, Prompt 1	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, Prompt 2	**83.46**	72.37	**67.66**	68.76	59.30	**78.43**	**61.63**	**69.60**
On FS, LoRA, Prompt 2	76.86	**73.57**	65.56	**69.06**	**59.69**	78.23	60.77	69.40
Average	80.16	72.97	66.61	68.91	59.49	78.33	61.20	69.50

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

2.1 Analysis of GRPO-based Reasoning versus SFT-based Direct Answering

We mainly compare the "SFT on Qwen2-Audio-7B-Instruct" and "GRPO on Qwen2-Audio-7B-Instruct" parts of Table 2. By observing the "Average" row, we find that SFT performs excellently on easy questions. In contrast, GRPO demonstrates superior performance on medium and hard questions. We believe this occurs because medium and hard questions require reasoning rather than relying solely on base knowledge for direct answers. Conversely, the weaker performance on easy questions may stem from redundant reasoning or process errors that propagate to the answer.

We conclude that GRPO is more effective for complex tasks that models cannot solve directly. However, its reasoning on simple tasks requires optimization to minimize redundancy.

2.2 Analysis of Explicit versus Implicit Prompting Strategies

We compare four LoRA experiments in the "GRPO on Qwen2-Audio-7B-Instruct" part. First, the two prompts produce distinct outputs. Unlike explicit prompts (Prompt 2), models using implicit prompts (Prompt 1) directly output answers similar to SFT, without explicit CoTs. Detailed prompt settings are provided in Appendix A.1.

Within the Prompt 1 experiments, models trained on FS consistently perform better. This suggests that implicit prompts induce SFT-like behavior reliant on data scale for generalization, whereas explicit reasoning enables deep understanding. For the two prompts, Prompt 2 outperforms Prompt 1 by approximately 0.15 on average. Specifically, Prompt 1 scores about 0.1 higher on easy and medium questions but lags by 1.1 on hard questions. These results confirm that reasoning on easier questions may lead to redundancy or error propagation, while harder questions require deeper reasoning.

Based on the findings, we conclude that while GRPO excels at complex tasks, its reasoning on simple tasks requires optimization to reduce redundancy. Specifically, the model should dynamically adapt reasoning length to question difficulty. By shortening reasoning for simple questions and extending it for complex ones, the model improves both performance and efficiency.

3 Enhancing LALMs with Difficulty-Adaptive Reasoning

The analysis above indicates that different questions require different reasoning lengths. Therefore, we aim to link question difficulty with reasoning length. This enables shorter reasoning for simple questions and deeper reasoning for difficult ones. Specifically, we define two difficulty-adaptive standards. The first is outcome-oriented and utilizes the group accuracy of rollout samples. The second is process-oriented and relies on the audio attention entropy of the current sample. We then apply these standards to a length based reward function to explicitly link question difficulty with reasoning length. The following is a detailed description of these two components.

3.1 Defining Model-perspective Difficulty

Here, we detail the two proposed difficulty-adaptive standards which we refer to as **G**roup **R**atio **D**ifficulty **R**eward (**GRDR**) and **G**roup **A**udio **A**ttention **D**ifficulty **R**eward (**GA²DR**). Both reuse variables already produced during training and therefore do not affect training speed.

GRDR. The first is outcome-oriented, determined by the proportion of correct samples within a rollout group. For a group size $G = 8$, we follow Liu et al. (2025) and classify questions as easy, medium, or hard based on correct response counts of at least 6, 3 to 5, and fewer than 3, respectively. We represent difficulty with γ, assigning values of 0, 0.5, and 1 to these categories. The specific formula is defined below.

$$C = \sum_{i=1}^{G} c_i \ (G = 8). \tag{1}$$

$$\gamma = \begin{cases} 0, & C \geq 6, \\ 0.5, & 3 \leq C < 6, \\ 1, & C < 3, \end{cases} \tag{2}$$

Here, $c_i \in \{0, 1\}$ denotes whether rollout sample o_i is correct, and C represents the total number of correct answers within the group.

GA²DR. The second is process-oriented and tailored to the audio modality. As demonstrated in Zhang et al. (2025), more dispersed attention corresponds to greater model uncertainty, while lower entropy reflects lower uncertainty. Accordingly, we treat high-entropy cases as more difficult questions and low-entropy cases as relatively simpler ones. Technically, we define $a_j^{(n)}$ as the attention weights from the last token in the final hidden layer over all previous positions. We average these across N heads to derive $\bar{p}_j$ for the audio tokens. We then compute the entropy and normalize it across the batch to $[0, 1]$ to define the difficulty level. The complete calculation is as follows:

$$a_j^{(n)} = \mathbf{A}_{T,j}^{(n)} \, , \, \bar{p}_j = \frac{1}{N} \sum_{n=1}^{N} a_j^{(n)}. \tag{3}$$

$$H = -\sum \bar{p}_j \log \bar{p}_j \, , \, (j \in \mathcal{M}). \tag{4}$$

$$\gamma^{(b)} = \frac{H^{(b)} - \min_{b' \in \mathcal{B}} H^{(b')}}{\max_{b' \in \mathcal{B}} H^{(b')} - \min_{b' \in \mathcal{B}} H^{(b')}}. \tag{5}$$

Here, $\mathbf{A}$ denotes the full attention matrix, T is the token count, and $\mathcal{M}$ represents the indices for audio attention. H denotes the computed entropy and $\mathcal{B}$ is the batch size. The final term $\gamma^{(b)}$ signifies the difficulty value of the b-th sample in the batch.

3.2 Difficulty-Adaptive Length-Based Reward

After defining question difficulty, we link it to reasoning length. Our core idea assigns short reasoning to simple questions and long reasoning to difficult ones using dynamic reward values. We implement a length based reward using a negative exponential function as shown in Figure 2. This function adjusts values based on the ratio between the current and model's maximum reasoning lengths. Incorrect samples receive a penalty that decreases as reasoning extends, which encourages further reasoning. Correct samples receive higher rewards as reasoning shortens, which encourages conciseness.

Specifically, GRDR (Plot a) employs six discrete difficulty curves while GA²DR (Plot b) allows the curve exponent to vary continuously. Both methods follow the same principle where simple questions correspond to steeper curves and difficult ones to flatter curves. This implies that easy questions permit shorter reasoning for a given reward value. Conversely, difficult questions receive higher rewards for the same length. This mechanism effectively promotes concise reasoning for simple tasks and deeper reasoning for difficult ones. We present the detailed calculation below.

$$r_i = sign(o_i) \cdot e^{-k(\gamma) l_{o_i}} \tag{6}$$

$$k(\gamma) = (1 - \gamma) k_{easy} + \gamma k_{hard} \tag{7}$$

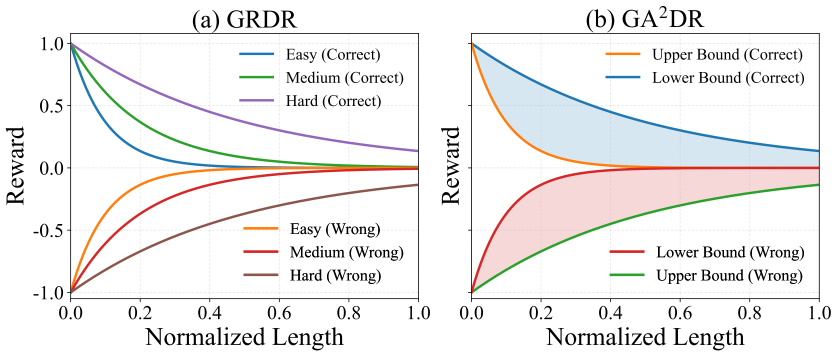


Figure 2: Curves of GRDR and GA^{2}DR with normalized length.

Here, the $sign$ function takes values of 1 or -1 to indicate whether a sample is correct. l_{o_i} represents the ratio of the sample's output length to the model's maximum completion length and ranges from $[0, 1]$. We derive adaptive difficulty factor $k(\gamma)$ via linear interpolation to map the defined difficulty standards onto optimization curves.

4 Experiments

Data. For training, we use FS , which is approximately twice the size of AVQA. The detailed data statistics have been presented in Section 2.

Setup. We conduct experiments using Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B, both fine-tuned via LoRA with Prompt 2. We perform the primary evaluations on MMAU-Test-Mini. To further verify generalization, we also assess performance on MMAU-Test-Mini(v05.15.25) and MMAR (Ma et al., 2025). The former features improved Q&A formulations and enhanced audio quality. The latter serves as a dataset for assessing deep reasoning capabilities across complex multitask and multilingual scenarios. All the experiments use ACC as the metric and Appendix A.1 provides additional details.

4.1 Main Results

Table 3 presents our main experiments on MMAU-Test-Mini. It includes results from four audio or omni models and our GRPO experiments on Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B, using both the basic Truncation Reward (TR) (Liu et al., 2025) and our proposed GRDR and GA^{2}DR. Besides, we extend MMAU-Test-Mini with model-perspective difficulty labels annotated by the four audio or omni models. Further results on MMAU-v0515 and MMAR are provided in Table 4 and Table 5. More details including the model-perspective annotations can be found in Appendix A.3.

Performance across Models. First, we analyze four audio and omni models. Their overall performance reflects different capability levels and can be seen as the average proficiency of representative LALMs, but clear differences emerge on hard questions. In particular, the leading Gemini2.5-Pro does not consistently outperform others and is even surpassed by some models. We argue that easy samples mainly rely on textual understanding, which is largely determined by LLM backbones. In contrast, hard samples require deep audio understanding, where LALMs can outperform omni models. This trend is evident between Qwen2.5-Omni-7B and Kimi-Audio-7B-Instruct. Sharing the same Qwen2.5 backbone and Whisper encoder, they perform similarly on easy samples, while the latter shows clear advantages on hard questions.

Second, we compare our proposed GRDR and GA^{2}DR with TR. For Qwen2-Audio-7B-Instruct, the methods show only marginal performance differences. This may result from the weaker base model performance with shorter initial output length. However, GRDR still achieves the best average score and performs relatively better on hard samples. For Qwen2.5-Omni-7B, both of our methods deliver gains, especially on medium and hard samples ($p \leq 0.05$). This demonstrates that our methods effectively utilize difficulty to assign appropriate rewards, and improve performance on challenging tasks.

Third, we compare GRDR and GA^{2}DR focusing on the results with Qwen2.5-Omni-7B. The two methods differ fundamentally as GRDR is outcome-oriented whereas GA^{2}DR is process-oriented. GA^{2}DR excels on medium and hard ones and comparable to GRDR on easy questions. We attribute this to their differing difficulty definitions.

Models	MMAU-Test-Mini									
	Sound	Music	Speech	Easy	Easy†	Medium	Medium†	Hard	Hard†	Avg
Compared 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*	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	**73.57**	65.56	**69.06**	59.69	**93.92**	78.23	59.81	**60.77**	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	72.45	66.66	**60.07**	93.16	80.00	58.87	59.91	32.81	70.20
+ GA^{2}DR	71.77	**74.25**	66.66	59.69	93.35	**81.37**‡	**61.21**‡	60.34	**33.20**‡	**70.90**

Table 3: Evaluations on MMAU-Test-Mini. Here, † denotes the model-perspective difficulty annotations, $*$ denotes a leading proprietary model and ‡ denotes the result significance $p \leq 0.05$ (compared to directly GRPO).

Models	MMAU-Test-Mini-v0515			
	Sound	Music	Speech	Avg
Compared 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^{2}DR	83.18	71.55	75.67	**76.80**

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

Models	MMAR			
	Sound	Music	Speech	Avg
Compared 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^{2}DR	64.84	54.85	65.30	**62.90**

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

GRDR uses only three discrete levels. In contrast, GA^{2}DR applies an unconstrained scale. This proves particularly beneficial for harder ones.

In summary, both GRDR and GA^{2}DR achieve clear gains, especially on hard questions, showing the effectiveness of our proposed methods.

Performance across Datasets. In terms of overall performance, our GA^{2}DR still achieves the best results on MMAU-v0515 and MMAR compared with all open-source baselines. This verifies the generalization ability of our methods. However, the performance of GRDR is weaker on MMAR. We attribute this difference to the nature of this approach. GRDR is outcome-oriented and highly sensitive to noise in rollout samples. As model capability improves, rollout accuracy tends to become high in later steps, which reduces the perceived difficulty gap between samples. As a result, most samples are eventually treated as easy, leading to overly short optimization and, in turn, reward hacking. Consequently, it performs reasonably well on the relatively simpler MMAU benchmarks but degrades significantly on the more challenging MMAR benchmark. In contrast, GA^{2}DR is not affected by such noise because its difficulty

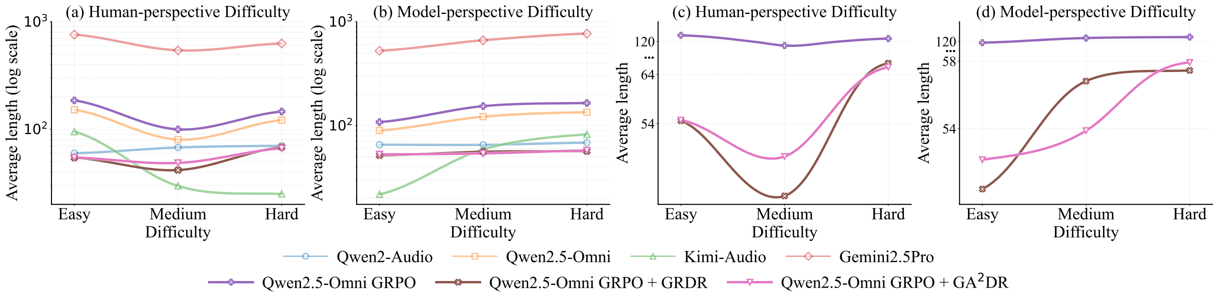


Figure 3: The trend of average reasoning length on MMAU-Test-Mini, under both human-perspective and model-perspective difficulty. Length is measured in tokens.

estimation is based on the audio attention characteristics of each question and is normalized within each batch. This preserves discrimination among samples, removes much of the randomness, and reduces the risk of reward hacking. As a result, it yields more stable performance across datasets and consistently strong results on more challenging tasks. We also conduct additional experiments to further verify generalization. More details are provided in Appendix A.4.

4.2 Analysis Regarding Reasoning Lengths

Here, we compare the output reasoning lengths of different models on MMAU-Test-Mini from both difficulty perspectives in Table 3. Due to large differences in output length across models, plots (a) and (b) show log-scaled length curves for all models, while plots (c) and (d) present the raw length curves for three Qwen2.5-Omni-7B-based models. Considering that our methods do not change the architecture of the model itself, the shorter output length means higher efficiency. All lengths are measured using the Qwen2 tokenizer. Detailed length statistics are available in Appendix A.5.

First, we compare reasoning length under two difficulty perspectives and find clear differences. From the human perspective, reasoning remains long for easy and hard ones, while from the model perspective, it grows with difficulty. This mismatch suggests that human-defined difficulty may conflict with the model during training. Moreover, models at different training stages behave differently. These findings support using model-perspective difficulty and updating it adaptively during training, which is exactly what our method provides.

Second, when comparing reasoning lengths of the two methods, we observe different patterns. From the human perspective, reasoning lengths are similar for easy and hard samples, while from the model perspective, both increase with difficulty. A clear gap emerges on medium-difficulty samples under both views, which we attribute to differences between GRDR and GA^{2}DR. In GA^{2}DR, entropy is normalized within each batch and mapped through linear interpolation, leading to a more uniform difficulty distribution and a smoother length curve than GRDR. Overall, both methods generate shorter reasoning across all difficulty levels, demonstrating their effectiveness.

In summary, our difficulty-adaptive length-based reward is effective, enabling concise reasoning for simple questions, deeper reasoning for difficult ones, and substantially helps LALMs reason smartly alongside higher performance.

Settings	**MMAU**	**MMAU v0515**	**MMAR**	**Avg**
GRPO	69.40	75.90	59.90	68.36
k=2	69.40	76.40	61.40	69.06
k=6	69.20	76.40	61.80	69.13
k=10	69.60	76.60	62.20	69.46
Ours	**70.90**	**76.80**	**62.90**	**70.20**

Table 6: Performance of different k settings across MMAU, MMAU-v0515 and MMAR.

4.3 Ablation Study on Difficulty Factor k

Here, we conduct an ablation study on our proposed difficulty-adaptive length-based reward. We remove the mechanism in Equation 6 and replace the adaptive difficulty factor $k(\gamma)$ in Equation 7 with fixed difficulty factor k. Then we experiment with various settings of k and Table 6 presents the average results across all three datasets, while Appendix A.6 provides more detailed results.

Regarding overall performance, all fixed k-value settings outperform the direct GRPO baseline,

Attention Layer	MMAU	MMAU v0515	MMAR	Avg
First	70.40	76.40	62.60	69.80
Mid	70.50	77.40	62.10	70.00
Last	70.90	76.80	62.90	**70.20**

Table 7: Performance of different attention layers for GA^{2}DR across MMAU, MMAU-v0515 and MMAR.

demonstrating the effectiveness of the length based reward itself. Average performance across the three datasets improves as k increases, since larger k leads to shorter outputs, indicating a need for length optimization. However, all fixed settings are inferior to the difficulty-adaptive reward, showing that adaptive difficulty is necessary. Using a single reward curve for all questions creates imbalance and degrades performance. Overall, these results confirm the effectiveness of the difficulty-adaptive length-based reward strategy.

4.4 Analysis Regarding Different Attention Layers for GA^{2}DR

Here we further analyze impact of different attention layers for GA^{2}DR. Specifically, we evaluate the first, the mid and the last attention layer across all three datasets in Table 7 and more detailed results are provided in Appendix A.6.

The overall performance across three benchmarks confirms that the last attention layer yields the best results, which justifies our selection. Additionally, performance consistently improves as layer depth increases. This trend aligns with our assumption that deeper layers contain more semantic information and better represent the model's final state during problem-solving. Therefore, selecting the last layer is the optimal choice for calculating the difficulty standards in GA^{2}DR.

5 Related work

Recent advances in LLMs have promoted the development of MLLMs. In the audio domain, LALMs like Qwen2-Audio(Chu et al., 2024), AudioFlamingo (Kong et al., 2024), and SALMONN (Tang et al., 2024) perform well on basic tasks but struggle with complex scenarios. More recent models, including Qwen2.5-Omni (Xu et al., 2025) and Kimi-Audio (KimiTeam et al., 2025), exhibit stronger ability but still remain constrained on challenging tasks. To address this, Audio-Reasoner (Xie et al., 2025) leverages large-scale CoT data for SFT, while R1-AQA (Li et al., 2025) and Omni-R1 (Rouditchenko et al., 2025) adopt reinforcement learning, to encourage self-driven reasoning without reliance on CoT annotations.

However, prior work primarily focuses on overall performance and often overlooks the distinctions between GRPO and SFT. We systematically analyze these aspects to clarify the mechanisms of different approaches. Our findings reveal that the explicit reasoning in GRPO is essential for difficult questions but often leads to redundancy on simple ones. Consequently, we propose optimizing reasoning length based on difficulty, minimizing redundancy for simple tasks while fostering deeper reasoning for complex ones.

Reasoning efficiency thus becomes a major challenge, with studies (Qu et al., 2025) identifying issues like redundancy and overthinking. While RL-based methods explore reward functions (Arora and Zanette, 2025; Aggarwal and Welleck, 2025; Shen et al., 2025), most rely on fixed length thresholds that ignore problem types and evolving model capabilities. Liu et al. (2025) introduces a difficulty-aware dynamic approach but needs an external dataset split. Similarly, in LALMs, Audio-Thinker (Wu et al., 2025) distinguishes when to reason but it does not further optimize the samples that require reasoning. To tackle these challenges, we propose two difficulty-adaptive length-based rewards to enhance reasoning efficiency in LALMs. By fostering conciseness for easy ones and depth for hard ones, and supporting adaptive optimization as the model evolves, our method reduces overall length while maintaining superior performance.

6 Conclusion

In this work, we study how LALMs can perform reasoning more effectively and efficiently. We first analyze two key questions: under what conditions SFT or GRPO is more effective, and whether performance gains arise from explicit reasoning. Our results show that explicit reasoning with GRPO is more effective, but reasoning length must be further optimized. Thus, we propose two difficulty-adaptive length-based rewards that explicitly link reasoning length to difficulty, encouraging concise reasoning for simple questions and deeper reasoning for more challenging ones. Extensive experiments on three datasets, together with comprehensive ablation studies, demonstrate the improved overall performance and higher efficiency.

Limitations

In this work, we focus on enabling LALMs to adjust reasoning length based on question difficulty and achieve notable gains in both performance and efficiency. However, some limitations remain. We do not explicitly constrain or optimize the quality of the CoT itself or its consistency with the final answer through reward design. Therefore, we conduct a qualitative analysis of the outputs produced by our two proposed rewards in Appendix A.9. Beyond confirming that our method does not cause the loss of critical reasoning steps, we also observe common and reasonable output patterns among strong reasoning models. In future work, we plan to introduce reward models that explicitly target CoT quality and its consistency with the final answer, aiming to produce higher-quality and more consistent reasoning outputs.

Ethics Statement

This work adheres to established ethical standards. All datasets are obtained and used in compliance with relevant usage guidelines, with no violations of privacy. We make consistent efforts to avoid bias or discriminatory outcomes throughout the research process. No personally identifiable information is involved. Elements in Figure 1 drawn from selected film or television works are used solely for academic presentation in the paper, not for any commercial or non-academic purposes, and strictly comply with fair use provisions under copyright law. All large language models are used only for text or image refinement and do not participate in the methods, code, experiments, or conclusions. No experiments raise privacy or security concerns. We remain committed to transparency and integrity throughout the research process.

References

Pranjal Aggarwal and Sean Welleck. 2025. L1: controlling how long A reasoning model thinks with reinforcement learning. *CoRR*, abs/2503.04697.

Daman Arora and Andrea Zanette. 2025. Training language models to reason efficiently. *CoRR*, abs/2502.04463.

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. 2024. Qwen2-audio technical report. *CoRR*, abs/2407.10759.

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, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Yuan Gong, Jin Yu, and James R. Glass. 2022. 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.

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, and 21 others. 2025. Kimi-audio technical report. *CoRR*, abs/2504.18425.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. 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.

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

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. 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.

Wei Liu, Ruochen Zhou, Yiyun Deng, Yuzhen Huang, Junteng Liu, Yuntian Deng, Yizhe Zhang, and Junxian He. 2025. Learn to reason efficiently with adaptive length-based reward shaping. *CoRR*, abs/2505.15612.

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, and 15 others. 2025. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *CoRR*, abs/2505.13032.

Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. 2017. 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*.

Ruotian Peng, Yi Ren, Zhouliang Yu, Weiyang Liu, and Yandong Wen. 2025. Simko: Simple pass@ k policy optimization. *arXiv preprint arXiv:2510.14807*.

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. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *CoRR*, abs/2503.21614.

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

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. 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*.

Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, and Shiguo Lian. 2025. DAST: difficulty-adaptive slow-thinking for large reasoning models. *CoRR*, abs/2503.04472.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. 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*.

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

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audioreasoner: Improving reasoning capability in large audio language models. *CoRR*, abs/2503.02318.

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. 2025. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.

Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. 2021. 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.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. 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.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. 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.

Zhisong Zhang, Yan Wang, Xinting Huang, Tianqing Fang, Hongming Zhang, Chenlong Deng, Shuaiyi Li, and Dong Yu. 2025. Attention entropy is a key factor: An analysis of parallel context encoding with full-attention-based pre-trained language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 9840–9855. Association for Computational Linguistics.

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. 2025. 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.

A Appendix

A.1 Setup

Models. We use Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B as base models, as they are widely adopted open-source LALMs and enable fair comparison with prior work. Training is performed with ms-swift (Zhao et al., 2025) on three A100-40G GPUs, using two for training and one for vLLM-based inference (Kwon et al., 2023). Additional details are provided in Appendix A.2.

Datasets. We use two training datasets, FS and AVQA (Yang et al., 2022). AVQA is a subset of FS, which further includes four additional datasets. For AVQA, we keep only audio–text pairs and replace "video" with "audio" in the questions. For analysis, we mainly evaluate on MMAU-Test-Mini (Sakshi et al., 2025), with AirBench Foundation-Sound (Yang et al., 2024) as a secondary reference. Main experiments are conducted on MMAU-Test-Mini, MMAU-v0515-Test-Mini, and MMAR.

GRPO. GRPO has been widely used in LLMs and MLLMs and has shown strong performance. Our implementation follows prior work (DeepSeek-AI et al., 2025; Li et al., 2025). Its key feature is computing the policy advantage using the average reward of grouped samples. Given an input question, sampled responses, and their rewards, the advantage is computed as follows:

$$A_i = \frac{r_i - mean(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}. \quad (8)$$

Here, A_i denotes the advantage used to optimize the policy model, $\{r_1, r_2, \cdots, r_G\}$ represents the set of reward values corresponding to each sampled output within the group $\{o_1, o_2, \cdots, o_G\}$, 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:

$$\mathcal{J}_{GRPO}(\theta) = \frac{1}{G}\sum_{i=1}^{G}(M_i - \beta\mathbb{D}_{KL}) \quad (9)$$

$$M_i = min(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i) \quad (10)$$

$$\rho_i = \frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{old}}(o_i \mid q)} \quad (11)$$

$$\mathbb{D}_{KL} = \frac{\pi_{ref}(o_i \mid q)}{\pi_\theta(o_i \mid q)} - log\frac{\pi_{ref}(o_i \mid q)}{\pi_\theta(o_i \mid q)} - 1 \quad (12)$$

Here, ϵ and β are hyper-parameters.

Models	MMAU-Test-Mini Sound	Music	Speech	Avg
All based on Qwen2-Audio-7B-Instruct				
R1-AQA	68.77	64.37	63.66	65.60
Ours (SFT)	63.06	56.59	57.66	59.10
+ GRPO (AVQA)	67.27	66.17	59.46	64.30
+ GRPO (FS)	70.57	65.57	66.07	**67.40**

Table 8: Evaluations on MMAU for cold-start methods.

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 4: Different prompt templates, where Prompt1 is the implicit prompt and Prompt2 is the explicit prompt.

Cold-start GRPO. A common approach for learning reasoning patterns is Cold-Start, which distills outputs from strong models and applies SFT before GRPO. We evaluate this by sampling data from the FS training set, distilling it with Gemini2.5-Pro, and using Qwen3-235B-A22B[3] to retain samples with correct and consistent reasoning and answers. We select 200 samples per task to form a 1,000-sample dataset for Cold-Start SFT on Qwen2-Audio-7B-Instruct. The model after two SFT epochs is then used to initialize GRPO. Results are shown in Table 8.

Implicit and Explicit Prompt. To test whether CoT benefits come from explicit outputs or implicit reasoning, we design two prompts shown in Figure 4. In both settings, the model outputs answers within <answer> </answer>. Prompt1 does not require CoT, while Prompt2 generates CoT within <think> </think>. A Format-Reward enforces this structure in Prompt2, whereas Prompt1 only constrains the answer format.

Truncation Reward. In addition to our proposed rewards, we include a simple length-threshold baseline. This method sets a threshold L_T and assigns positive rewards to correct rollouts with length no greater than L_T, while longer rollouts are penalized

[3]https://huggingface.co/Qwen/Qwen3-235B-A22B

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-based Reward	
Truncation Reward L_T	120 / 400
Truncation Reward σ	-0.5
GRDR and GA^{2}DR l_{min}	0.1
GRDR and GA^{2}DR k_{hard}	2
GRDR and GA^{2}DR k_{easy}	10

Table 9: Hyper-parameters for SFT and GRPO, including different settings under LoRA and Full.

even if correct. The reward is computed as follows:

$$r_i = \begin{cases} 1, & \text{if } L_{o_i} \le L_T \textit{ and } o_i \textit{ is correct}, \\ \sigma, & \text{if } L_{o_i} > L_T. \end{cases} \tag{13}$$

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

A.2 Hyper-parameters

Here, we provide explanations of the hyperparameter settings in Table 9, including the two proposed rewards and the basic TR hyperparameter settings.

A.3 Model-perspective Difficulty On MMAU

In Sakshi et al. (2025), question difficulty is manually annotated by experts, yielding high-quality but costly and fixed labels. Here, we introduce model-perspective difficulty to reflect average model capability. We run four models, Qwen2-Audio-7B-Instruct, Qwen2.5-Omni-7B, Kimi-Audio-7B-Instruct[4], and Gemini2.5-Pro-0506, on MMAU-Test-Mini with the same random seed, and define difficulty by the number of models answering each question correctly. The label distribution is shown in Table 10.

[4] https://huggingface.co/moonshotai/Kimi-Audio-7B-Instruct

Orig Diff.	Total-Num Orig	Total-Num New	Num Un-Chg	Num Chg	New Diff.	Chg Num
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

Table 10: 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.

A.4 Impact of Threshold Ratio on Our Proposed Reward

Since the negative exponential approaches 1 as its exponent tends to 0, it can be viewed as implicit reasoning, though this may harm readability. To analyze this effect, we introduce a minimum threshold ratio l_{min}. Samples with relative length below this threshold are set to 1, while longer samples are further normalized. The parameter ζ then applies this secondary normalization.

$$r_i = sign(o_i) \cdot e^{-k(\gamma)\zeta(l_{o_i};l_{min})} \tag{14}$$

$$\zeta(l_{o_i}; l_{min}) = max(0, \frac{l_{o_i} - l_{min}}{1 - l_{min}}) \tag{15}$$

Table 11, Table 12, and Table 13 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.

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. We emphasize the former two benchmarks, as MMAU-Test-Mini contains lower quality Q&As and audio. Overall, introducing a length threshold helps counter the reward curve's tendency to drive completion length toward zero, preventing excessive shortening that would weaken CoT reasoning and reduce performance.

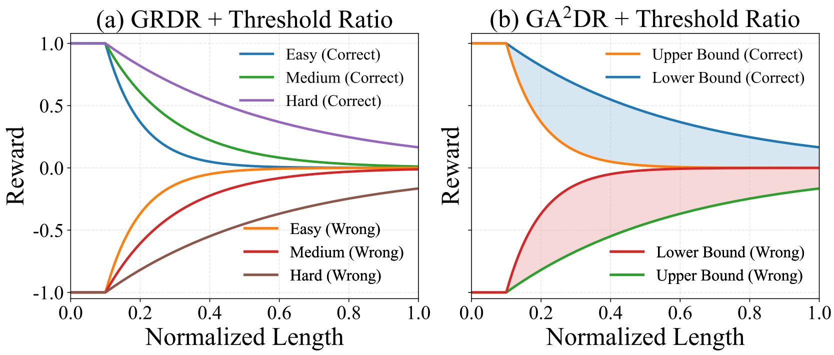


Figure 5: Curves of GRDR and GA^{2}DR with normalized length and the threshold ratio l_{min}.

Models	MMAU-Test-Mini									
	Sound	Music	Speech	Easy	Easy†	Medium	Medium†	Hard	Hard†	Avg
Based On Qwen2.5-Omni-7B										
GRDR	71.47	72.45	66.66	**60.07**	93.16	80.00	58.87	59.91	32.81	70.20
+ $l_{min} = 0.1$	**72.07**	69.76	63.36	55.81	**93.92**	79.60	55.14	57.75	27.41	68.40
GA^{2}DR	71.77	**74.25**	**66.66**	59.69	93.35	**81.37**	**61.21**	**60.34**	**33.20**	**70.90**
+ $l_{min} = 0.1$	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 GRDR, GA^{2}DR, and their variants with added length threshold ratios on MMAU-Test-Mini. Here, unmarked models correspond to $l_{min} = 0$, and † denotes the model-perspective difficulty annotations.

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

Table 12: The performance of our proposed rewards, and their variants with length threshold ratios. 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	**65.98**	63.00
GA^{2}DR	64.84	**54.85**	65.30	62.90
+ $l_{min} = 0.1$	**64.84**	53.39	63.60	**63.00**

Table 13: The performance of our proposed rewards, and their variants with length threshold ratios. Here, unmarked models correspond to $l_{min} = 0$.

When comparing GRDR and GA^{2}DR, the latter performs better on two benchmarks. This advantage comes from defining difficulty at the batch level, which aligns with batch-wise optimization. Prior work (Peng et al., 2025) shows that reinforcement learning can improve pass@1 but hurt pass@k due to unstable rollouts. GA^{2}DR reduces this issue by grouping rollouts with the same question and audio, thereby lowering variance.

A.5 Detailed length statistics of different models on MMAU

This section reports raw reasoning-length statistics for all models in our MMAU-Test-Mini experiments in Table 14. Lengths are measured with the Qwen2 tokenizer. Qwen2-Audio and Qwen2.5-Omni share this tokenizer, while it is also used for Gemini2.5-Pro due to its closed-source tokenizer. Although this may introduce minor deviation, Gemini2.5-Pro produces much longer reasoning, so relative differences remain clear.

A.6 Detailed Results on three benchmarks for Section 4.4 and Section 4.3

In Section 4.4 and Section 4.3, we analyze the choice of attention layers for GA^{2}DR and conduct ablation studies on the difficulty adaptive rewards. In this section, we present detailed results for each

Models	MMAU-Test-Mini					
	Easy	Easy†	Medium	Medium†	Hard	Hard†
Compared 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^{2}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

Table 14: The output length statistics on MMAU-test-mini, reported under both the human-perspective and model-perspective difficulty annotations. Here, † denotes the model-perspective difficulty annotations.

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^{2}DR	100	56.9
+ GA^{2}DR ($l_{min} = 0.1$)	95	94.2

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

benchmark. Tables 16, 17, and 18 report ablations on the k settings, while Tables 19, 20, and 21 present the analysis of attention layers.

A.7 Reasoning length vs. Accuracy

Here, we examine the relationship between output length and accuracy for GRDR and GA^{2}DR on MMAU-Test-Mini. From Figure 6 to Figure 9, samples are grouped into length intervals of comparable size, and correctly answered samples within each interval are further separated by difficulty, shown with progressively darker shades of blue.

A.8 Training Curves

This section presents the training curves of GRDR and GA^{2}DR on Qwen2.5-Omni, including gradient norm, KL divergence, and reward trends, shown in Figures 10 and 11. Both methods converge stably, without extreme gradients, abnormal KL divergence, or numerical spikes.

A.9 Case Study: Qualitative Analysis of Output Paradigms across Models

We analyze 20 MMAU-Test-Mini questions on Dissonant Emotion Interpretation with a focus on sarcasm cause detection, a task well suited for qualitative study due to complex multi-speaker audio. Detailed results are provided in Table 15.

For Qwen2-Audio-7B-Instruct, GRPO and GRPO with GRDR match the performance of Audio-Reasoner while producing much shorter outputs. In contrast, Cold-Start GRPO yields similar reasoning length but the worst performance, suggesting that weaker base models tend to mimic surface structures rather than learn effective rea-

Models	MMAU-Test-Mini									
	Sound	Music	Speech	Easy	Easy†	Medium	Medium†	Hard	Hard†	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 16: The performance of models trained with the single length based reward for different k settings on MMAU. Here, † denotes the model-perspective difficulty annotations.

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 17: The performance of models trained with the single length based reward for different k settings on MMAU-v0515.

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

Table 18: The performance of models trained with the single length based reward for different k settings on MMAR.

Layer	MMAU-Test-Mini									
	Sound	Music	Speech	Easy	Easy†	Medium	Medium†	Hard	Hard†	Avg
GRPO Based On Qwen2.5-Omni-7B (With length based Reward Only)										
First	72.07	71.55	67.56	60.85	94.30	79.60	57.00	60.77	32.81	70.40
Mid	72.97	71.25	67.26	58.14	94.87	80.78	60.74	61.63	28.95	70.50
Last	71.77	74.25	66.66	59.69	93.35	81.37	61.21	60.34	33.20	70.90

Table 19: The performance of GA^{2}DR with different attention layer selections on MMAU. Here, † denotes the model-perspective difficulty annotations.

Layer	MMAU-Test-Mini-v0515			
	Sound	Music	Speech	Avg
GRPO Based On Qwen2.5-Omni-7B				
First	84.08	70.35	74.77	76.40
Mid	81.38	73.35	77.47	77.40
Last	83.18	71.55	75.67	76.80

Table 20: The performance of GA^{2}DR with different attention layer selections on MMAU-v0515.

Layer	MMAR			
	Sound	Music	Speech	Avg
GRPO Based On Qwen2.5-Omni-7B				
First	62.42	55.34	64.28	62.60
Mid	61.81	53.39	63.94	62.10
Last	64.84	54.85	65.30	62.90

Table 21: The performance of GA^{2}DR with different attention layer selections on MMAR.

soning, making Cold-Start ineffective or harmful. For Qwen2.5-Omni-7B, performance differences are small since the base model already supports complex reasoning. Even so, our proposed rewards maintain accuracy while substantially shortening reasoning, leading to clear efficiency gains.

Qualitative inspection shows that strong models follow a consistent process of grounding audio cues, conducting step-by-step analysis, and producing a final answer. Weaker models often fail at early information extraction, causing errors to propagate. Figures 12–14 show outputs from six models, including Gemini2.5-Pro, Audio-Reasoner, and our four models. Correct options are highlighted in green, logical steps in blue, and redundancies in red.

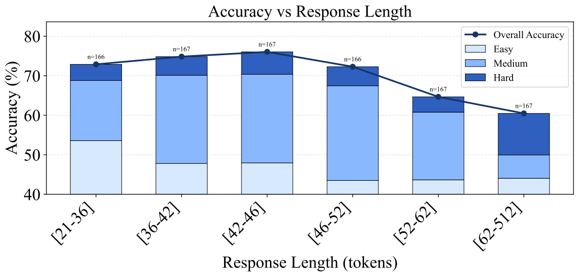


Figure 6: The trend of length and accuracy for the GRDR with human-perspective difficulty.

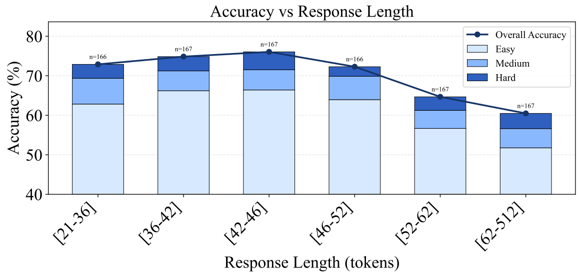


Figure 7: The trend of length and accuracy for the GRDR with model-perspective difficulty.

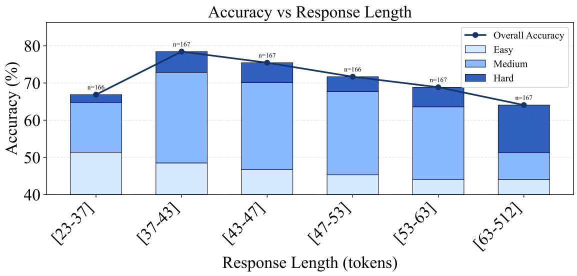


Figure 8: The trend of length and accuracy for the GA^{2}DR with human-perspective difficulty.

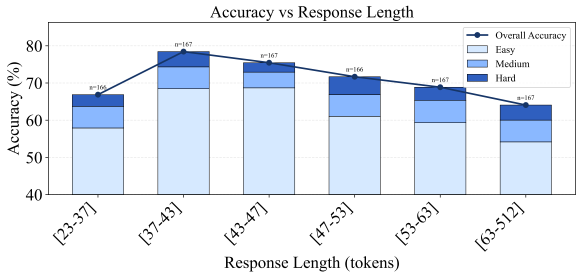
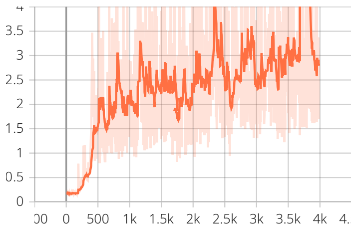
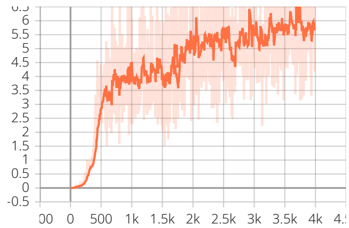


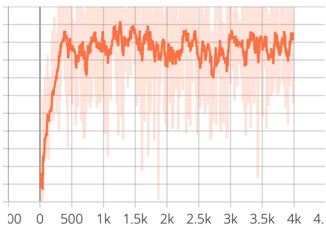
Figure 9: The trend of length and accuracy for the GA^{2}DR with model-perspective difficulty.



(a) Grad Norm

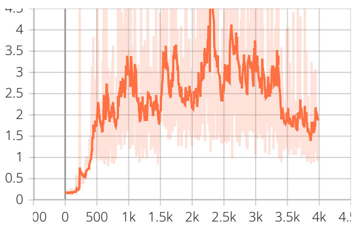


(b) KL Divergence

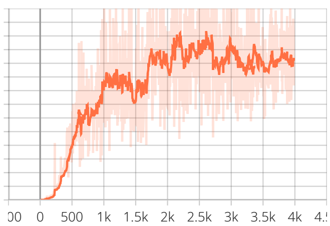


(c) Reward

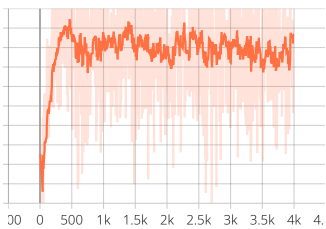
Figure 10: Training curves of gradient norm, KL divergence, and reward on GRDR.



(a) Grad Norm



(b) KL Divergence



(c) Reward

Figure 11: Training curves of gradient norm, KL divergence, and reward on GA^{2}DR.

Overall, effective reasoning relies on accurate grounding, structured analysis, and clear conclusions. Accordingly, weaker models should first focus on grounding tasks, stronger but less readable models may benefit from Cold-Start, and models with established structures can directly apply RL with our proposed rewards to improve both performance and efficiency.

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

<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.

Gemini2.5-Pro

<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.

Audio Reasoner

<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 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-Audio Cold-Start 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

<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.

Qwen2.5-Omni GRPO-GA^{2}DR

Figure 12: 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.

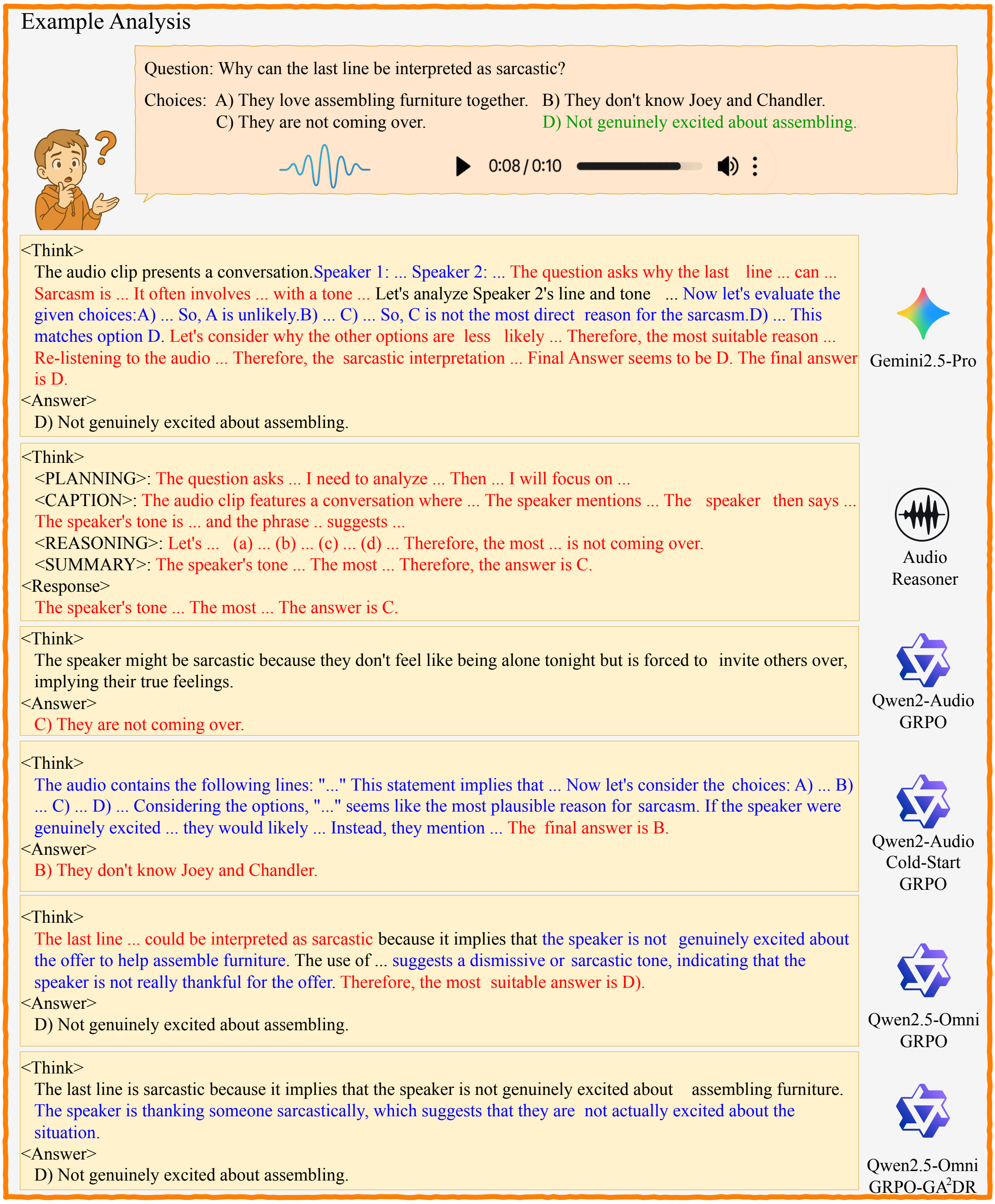


Figure 13: Output examples from six different models on a sarcasm cause detection question. The green part in the question indicates the correct option. The blue parts in the model outputs represent core content, while the red parts indicate redundancy.

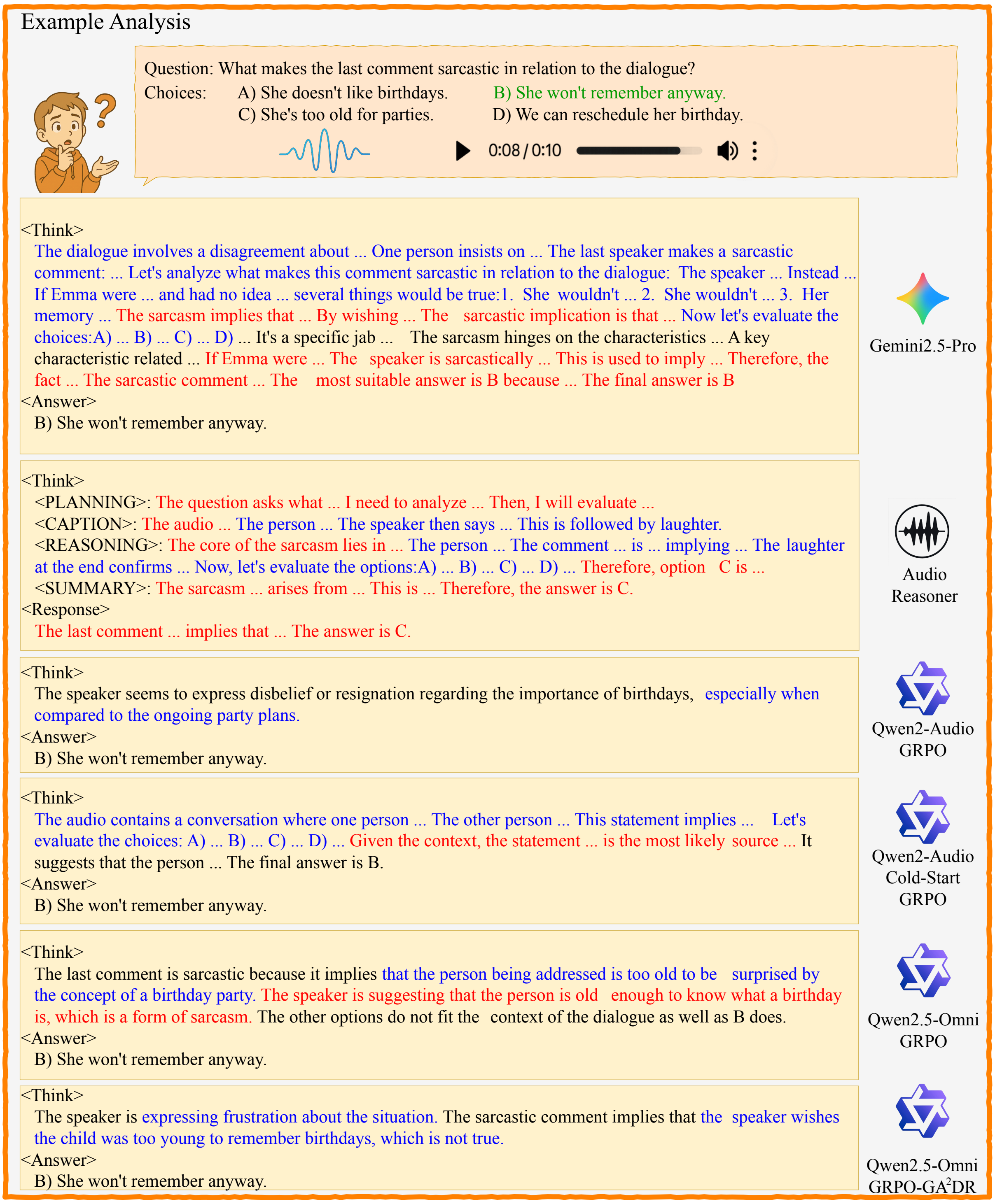


Figure 14: Output examples from six different models on a sarcasm cause detection question. The green part in the question indicates the correct option. The blue parts in the model outputs represent core content, while the red parts indicate redundancy.