

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 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

054 model on whether reasoning is necessary. However, for the samples that still perform reasoning, it
 055 lacks a fine-grained mechanism to modulate how much to reason. Therefore, an effective and efficient
 056 LALM should reason smartly by adapting its reasoning depth to the problem’s complexity, achieving
 057 short reasoning for simple questions and deeper reasoning for difficult ones. Based on this analysis
 058 and prior work, we propose a new length-based reward function that no longer depends on fixed
 059 thresholds. In addition, we introduce two difficulty-adaptive standards to complement this reward,
 060 enabling reasoning length to align more appropriately with question difficulty. Both approaches
 061 achieve strong performance on the MMAU benchmark (Sakshi et al., 2025), particularly on harder
 062 questions, while also producing much shorter reasoning than direct GRPO models, thereby greatly
 063 improving reasoning efficiency. Furthermore, we conduct a qualitative case study on models from
 064 the main experiments, providing a detailed analysis of the reasoning structures in their outputs.
 065

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
Sound Question Answering	TUT2017 (Mesaros et al., 2017)	3,744
	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-Instruct¹ and Qwen2.5-Omni-7B². More detailed experimental settings, including the two prompts, are provided in Appendix

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

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

108 A.2. At the same time, for those experiments that share common settings, we also add an additional
 109 group average in Table 2 to facilitate comparisons across groups.
 110

111 2.1 HOW DOES GRPO REASONING DIFFER FROM DIRECT ANSWERS WITH SFT?
 112

113 Here we mainly compare the “SFT on Qwen2-Audio-7B-Instruct” part and the “GRPO on Qwen2-
 114 Audio-7B-Instruct” part of Table 2. From the “Average” of the SFT part and the “Average” of
 115 the Prompt2 part, we observe that SFT performs very well on easy-level questions, while GRPO
 116 shows greater advantages on medium and hard questions. We believe this is mainly because, when
 117 facing medium and harder questions, the model cannot rely only on the base knowledge learned
 118 during pretraining to give direct answers and instead needs to learn how to use this knowledge
 119 through reasoning. In contrast, its weaker performance on easy questions is largely due to redundant
 120 reasoning content or errors made during the reasoning process that are carried over into the final
 121 answer, leading to incorrect results.
 122

123 Therefore, for the question in this subsection, we conclude that GRPO is more effective on com-
 124 plex tasks that models cannot solve directly, but its reasoning on simple ones still needs further
 125 optimization to reduce redundancy and potential error propagation.
 126

127 Table 2: The performance of different models under different training paradigms, fine-tuning strate-
 128 gies, training datasets, and prompting styles.
 129

130 Models	Airbench-Foundation			MMAU-Test-Mini				Avg
	131 Sound	132 Sound	133 Music	134 Speech	135 Easy	136 Medium	137 Hard	
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	83.46	72.37	67.66	68.76	59.30	78.43	61.63	69.60
On FS, LoRA, Prompt2	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

151
 152 2.2 DOES PERFORMANCE COME FROM EXPLICIT REASONING OR IMPLICIT ACTIVATION?
 153

154 This part mainly compares the four LoRA experiments under the “GRPO on Qwen2-Audio-7B-
 155 Instruct” part. First, it should be noted that the two prompts produce outputs with clear differences.
 156 Models trained with the implicit prompt do not generate a reasoning process for nearly every sample
 157 as those trained with the explicit prompt do; instead, they often directly produce answers in a way
 158 similar to SFT. When comparing results within the “Prompt1” (implicit prompt) experiments, mod-
 159 els trained on the larger FS dataset consistently achieve better performance. These results suggest
 160 that using implicit prompts introduces SFT-like characteristics, relying on larger datasets to achieve
 161 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”

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

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

173 3 ENHANCING LALMs WITH DIFFICULTY-AWARE ADAPTIVE REASONING

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

183 3.1 DEFINING MODEL-PERSPECTIVE DIFFICULTY

185 As mentioned above, our core idea is to encourage models to reason more on difficult questions and
 186 less on simple ones. Thus, our main approach is to dynamically adjust the reward based on both
 187 question difficulty and reasoning length. At the same time, since there are gaps between different
 188 definitions of difficulty perspectives, the model’s own perspective during training can better reflect
 189 the actual situation. Therefore, in this subsection, we provide a detailed explanation of the two pro-
 190 posed model-perspective difficulty-adaptive standards. We refer to these as **Group Ratio Difficulty**
 191 **Reward (GRDR)** and **Group Audio Attention Difficulty Reward (GA²DR)**.

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

$$198 \quad \gamma = \begin{cases} 0, & C \geq 6, \\ 200 \quad 0.5, & 3 \leq C < 6, \quad (G = 8), \quad C = \sum_{i=1}^G c_i \\ 201 \quad 1, & C < 3, \end{cases} \quad (1)$$

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

205 **GA²DR.** The second approach is more characteristic of the audio modality. As mentioned earlier,
 206 when the model’s attention to the audio segment is more dispersed, it suggests that the audio is
 207 complex and that the model struggles to identify a key point to solve the problem. In this case, the
 208 attention entropy is relatively large, corresponding to higher difficulty. Specifically, we use $a_j^{(n)}$,
 209 the attention after softmax from the last token position in the final hidden layer over all previous
 210 positions, and \bar{p}_j , the attention values assigned to audio tokens averaged across attention heads N ,
 211 and then compute the entropy. This entropy is normalized across the batch to a value in [0,1], which
 212 directly represents difficulty. The complete calculation process is as follows:

$$214 \quad a_j^{(n)} = \mathbf{A}_{T,j}^{(n)}, \quad \bar{p}_j = \frac{1}{N} \sum_{n=1}^N a_j^{(n)}, \quad H = - \sum \bar{p}_j \log \bar{p}_j, \quad (j \in \mathcal{M}). \quad (2)$$

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

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

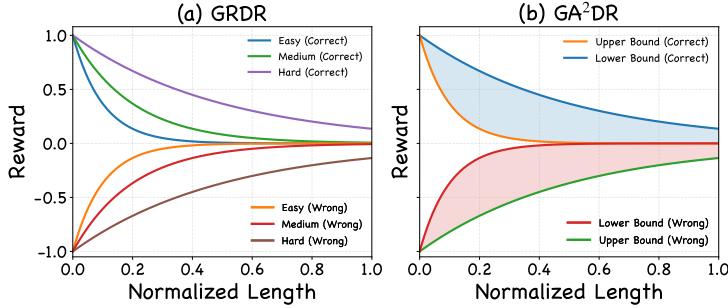


Figure 1: Curves of GRDR and GA²DR with normalized length.

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:

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

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

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.

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

278 Table 3: The performance of models on MMAU-test-mini with different base models and reward
 279 settings, reported under both the human-perspective difficulty annotations and the assigned model-
 280 perspective difficulty annotations. Here, \dagger denotes the model-perspective difficulty annotations and
 281 * 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 [*]	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	71.47	71.85	66.66	57.75	92.78	80.58	59.81	60.34	32.04	70.00

300
 301
 302 Table 4: Evaluations on MMAU-v0515. Here, * denotes a leading proprietary model.

303
 304 Table 5: Evaluations on MMAR. Here, * denotes a leading proprietary model.

Models	MMAU-Test-Mini-v0515				MMAR				
	Sound	Music	Speech	Avg	Sound	Music	Speech	Avg	
Baseline Models									
Qwen2-Audio	62.16	62.27	55.55	60.00	Qwen2-Audio	33.33	24.27	32.31	30.00
Qwen2.5-Omni	74.17	65.26	61.56	67.00	Qwen2.5-Omni	58.79	40.78	59.86	56.70
Kimi-Audio	78.97	60.47	66.96	68.80	Kimi-Audio	57.57	45.63	63.26	59.00
Gemini2.5-Pro*	76.57	73.95	80.78	77.10	Gemini2.5-Pro*	73.33	64.07	88.77	80.50
Based On Qwen2.5-Omni-7B									
GRPO	84.08	69.46	74.17	75.90	GRPO	60.00	48.05	62.24	59.90
+ TR	83.78	70.65	74.47	76.30	+ TR	64.84	49.51	63.94	61.90
+ GRDR	83.48	70.35	75.97	76.60	+ GRDR	61.21	51.94	65.30	61.20
+ GA ² DR	83.18	71.55	75.67	76.80	+ GA ² DR	64.84	54.85	65.30	62.90

318
 319
 320 4.1 MAIN RESULTS

321 Table 3 mainly presents three parts: the performance of four baseline models and our proposed
 322 rewards on Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B. We further report results on MMAU-
 323

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

324 v0515 and MMAR, together with the base model on these benchmarks, shown in Table 4 and Table
 325 5. TR denotes the basic Truncation Reward (Liu et al., 2025), with its formula provided in Appendix
 326 A.2. In addition, we extend MMAU-test-mini with model-perspective difficulty labels annotated by
 327 the four baseline models in Table 3, with details given in Appendix A.3.
 328
 329

330 **Performance across Models.** First, we analyze the four baseline models. Their overall perfor-
 331 mance reflects different capability levels and the typical range of most systems. However, they
 332 differ markedly on hard questions. Even Gemini2.5-Pro, which is nearly the strongest on average,
 333 performs poorly on hard items, while Qwen2-Audio, the weakest on average, shows a clear ad-
 334 vantage. We attribute this to the combined effects of the LLM backbone strength and the number
 335 of supported modalities. For easy questions, the LLM backbone is dominant because these items
 336 require minimal audio understanding and rely mainly on text comprehension. In contrast, hard
 337 questions demand stronger audio interpretation, where pure LALMs often outperform Omni-style
 338 models that integrate more modalities. This pattern is evident in Qwen2.5-Omni and Kimi-Audio,
 339 which share a “Whisper + Qwen2.5” architecture but still exhibit a noticeable gap on hard questions.
 340

341 Second, we compare our proposed GRDR and GA²DR with TR. For Qwen2-Audio-7B-Instruct,
 342 the three methods achieve similar overall performance, with GRDR performing best and TR worst,
 343 confirming the effectiveness of our approach. TR shows slightly better performance on medium
 344 questions, but GRDR is clearly stronger on hard ones. For Qwen2.5-Omni-7B, both of our meth-
 345 ods deliver clear overall gains. They substantially outperform TR on medium and hard questions,
 346 while maintaining comparable results on easy ones. This shows that our methods effectively uti-
 347 lize question difficulty to assign appropriate rewards, promote deeper reasoning on hard questions,
 348 and improve performance on challenging tasks. Unlike TR, which uses a fixed truncation length
 349 and only constrains part of the samples, our reward designs offer more balanced treatment across
 350 difficulty levels, leading to stronger overall results.
 351

352 Third, we compare GRDR and GA²DR, focusing on the results based on Qwen2.5-Omni-7B. The
 353 two methods differ in that GRDR is outcome-oriented, whereas GA²DR is process-oriented. GRDR
 354 performs better on easy questions, GA²DR is stronger on medium ones, and their performance is
 355 similar on hard questions. We believe this is due to their different difficulty definitions: GRDR
 356 uses only three levels, while GA²DR applies a finer, unconstrained difficulty scale, which partic-
 357 ularly benefits medium questions. For easy and hard questions, GA²DR is slightly weaker because
 358 normalization may place samples with similar audio attention entropy into different difficulty bins,
 359 diminishing its advantage.
 360

361 In summary, both GRDR and GA²DR achieve clear performance gains, especially on hard questions,
 362 showing that our methods can effectively adapt to different questions according to their difficulty.
 363

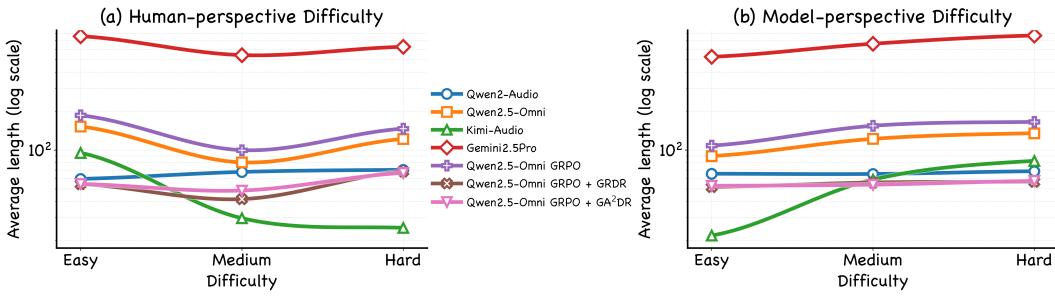
364 **Performance across Benchmarks.** To further evaluate our model in broader and more reasoning-
 365 intensive scenarios, we conduct additional evaluations on MMAU-v0515 and MMAR, which also
 366 help verify the generalization ability of our methods. In terms of overall performance, our two
 367 methods still achieve the best results on MMAU-v0515 compared with all baselines. On MMAR,
 368 however, the two methods behave differently: GA²DR maintains a clear performance lead, whereas
 369 GRDR falls behind the TR method.
 370

371 This difference can be explained by the nature of the two approaches. GRDR is outcome-oriented,
 372 while GA²DR is process-oriented. GRDR is highly susceptible to noise in rollout samples, making
 373 it prone to reward hacking. As a result, it performs reasonably well on the relatively simpler MMAU
 374 benchmarks but degrades significantly on the more challenging MMAR benchmark. In contrast,
 375 GA²DR is unaffected by such noise because it determines difficulty solely based on the model’s
 376 current behavior and the audio characteristics of each question. This removes much of the ran-
 377 domness and reduces the likelihood of reward hacking, leading to more stable performance across
 378 benchmarks and consistently strong results on more challenging tasks. We also conduct additional
 379 experiments to further verify generalization. The results confirm that GRDR is indeed affected by
 380 random noise introduced during rollout; however, once this issue is mitigated through appropriate
 381 constraints, its performance still surpasses all baseline methods. Detailed results and analysis are
 382 provided in Appendix A.4.
 383

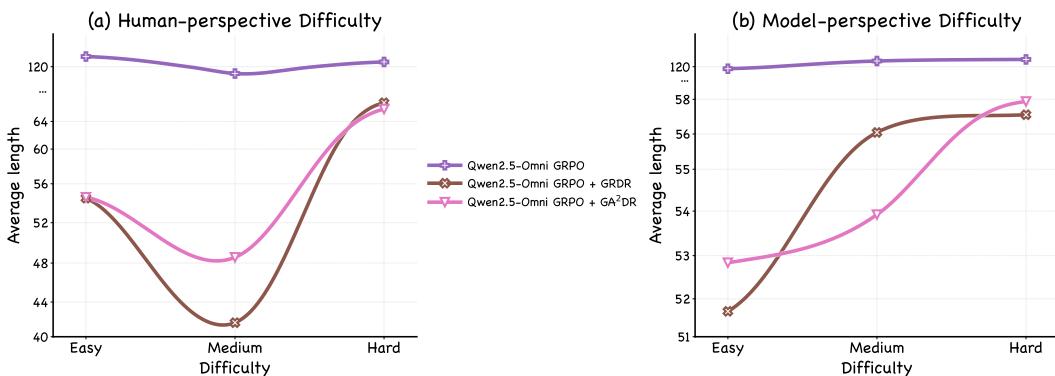
378 4.2 ANALYSIS OF REASONING LENGTH ACROSS MODELS
379

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

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



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



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

427 Second, when comparing reasoning lengths of our proposed rewards with other models, we find
428 that our two rewards produce shorter reasoning across all difficulty levels than most other models.
429 Under the human perspective, the lengths of our methods are close to direct GRPO for easy and hard
430 questions but are much shorter on medium questions. Under the model perspective, the curves of
431 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 GA²DR, 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

486 but remain limited on complex tasks due to short outputs and lack of reasoning. Later models like
 487 Qwen2.5-Omni (Xu et al., 2025) and Kimi-Audio (KimiTeam et al., 2025) demonstrate some initial
 488 reasoning ability but still rely heavily on SFT, leaving outputs fixed and dependent on pretraining
 489 data, thus performing poorly in complex scenarios. To address this, some studies extend SFT, such
 490 as Audio-Reasoner (Xie et al., 2025), which uses large-scale CoT-annotated pairs to achieve reasoning
 491 via SFT. Others adopt RL, as in R1-AQA (Li et al., 2025) and Omni-R1 (Rouditchenko et al.,
 492 2025), which avoids using data with CoT, and promotes self-driven reasoning.

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

501 Reasoning efficiency thus becomes a major challenge. Prior studies (Qu et al., 2025) highlight
 502 inefficiencies such as redundant content and overthinking. RL-based methods explore reward func-
 503 tion designs to address this issue (Arora & Zanette, 2025; Aggarwal & Welleck, 2025; Shen et al.,
 504 2025). Most methods rely on setting a fixed length threshold for optimization. On the one hand,
 505 such approaches do not link reasoning length with problem type and therefore cannot adapt to all
 506 kinds of problems. On the other hand, the fixed threshold cannot change with the model’s evolving
 507 ability during training. Liu et al. (2025) summarizes these efforts and proposes a difficulty-aware
 508 dynamic approach that improves both performance and efficiency. However, this method depends
 509 on an additional independent dataset during training to measure the model’s capability, which not
 510 only increases computational cost but also makes the results sensitive to how this dataset is selected.
 511 In LALMs, Audio-Thinker (Wu et al., 2025) instead tackles the problem through a “when to think”
 512 mechanism, dividing tasks into those requiring reasoning and those that do not, but it does not fur-
 513 ther optimize the samples that require reasoning. In this work, we propose two difficulty-adaptive
 514 length-based rewards to enable efficient and effective reasoning in LALMs. Our method maintains
 515 comparable or even superior performance while reducing overall reasoning length, encouraging con-
 516cise reasoning for simple questions, deeper reasoning for difficult ones, and progressive optimization
 517 as the model’s capability improves.

518 6 LIMITATIONS AND FUTURE WORK

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

527 7 CONCLUSION

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

540 ETHICS STATEMENT
541

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

548
549 REPRODUCIBILITY STATEMENT
550

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

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

561
562 REFERENCES
563

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

566
567 Daman Arora and Andrea Zanette. Training language models to reason efficiently. *CoRR*,
568 abs/2502.04463, 2025.

569
570 Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and
571 Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale
572 audio-language models. *CoRR*, abs/2311.07919, 2023.

573
574 Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv,
575 Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *CoRR*,
576 abs/2407.10759, 2024.

577 DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,
578 Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu,
579 Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao
580 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,
581 Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao,
582 Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,
583 Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang
584 Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai
585 Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang,
586 Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang,
587 Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang,
588 Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang,
589 R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye,
590 Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing
591 reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.

592
593 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.

594 Yuan Gong, Jin Yu, and James R. Glass. Vocalsound: A dataset for improving human vocal sounds
 595 recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing,*
 596 *ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022.

597

598 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-
 599 trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint*
 600 *arXiv:2410.21276*, 2024.

601 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec
 602 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard
 603 Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett,
 604 Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey
 605 Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben
 606 Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botaao Hao, Bowen
 607 Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin,
 608 Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris
 609 Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kap-
 610 pler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson,
 611 Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Eliz-
 612 abeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang,
 613 Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred
 614 von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace
 615 Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart An-
 616 drin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan,
 617 Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya. Openai o1 system card.
CoRR, abs/2412.16720, 2024.

618

619 KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen,
 620 Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao
 621 Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu,
 622 Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng
 623 Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan,
 624 Yutong Zhang, and Zaida Zhou. Kimi-audio technical report. *CoRR*, abs/2504.18425, 2025.

625

626 Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio
 627 flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Forty-
 628 first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27,
 629 2024*, 2024.

630

631 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph
 632 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model
 633 serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Prin-
 634 ciples, SOSP 2023, Koblenz, Germany, October 23-26, 2023*. ACM, 2023.

635

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

639

640 Samuel Lipping, Parthasarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-
 641 aqa: A crowdsourced dataset for audio question answering. In *30th European Signal Processing
 642 Conference, EUSIPCO 2022, Belgrade, Serbia, August 29 - Sept. 2, 2022*. IEEE, 2022.

643

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

647

648 Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen,
 649 Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan
 650 Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yux-
 651 uan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou,
 652 Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Chng Eng Siong, and Xie Chen. MMAR:

648 A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *CoRR*,
 649 [abs/2505.13032](https://arxiv.org/abs/2505.13032), 2025.

650

651 Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel
 652 Vincent, Bhiksha Raj, and Tuomas Virtanen. DCASE2017 challenge setup: Tasks, datasets and
 653 baseline system. In *Proceedings of the Workshop on Detection and Classification of Acoustic
 654 Scenes and Events, DCASE 2017, Munich, Germany, November 16-17, 2017*, 2017.

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

657

658 Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong
 659 Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-
 660 Sheng Hua, Bowen Zhou, and Yu Cheng. A survey of efficient reasoning for large reasoning
 661 models: Language, multimodality, and beyond. *CoRR*, [abs/2503.21614](https://arxiv.org/abs/2503.21614), 2025.

662 Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne,
 663 Rogério Feris, and James R. Glass. Omni-r1: Do you really need audio to fine-tune your au-
 664 dio llm? *CoRR*, [abs/2505.09439](https://arxiv.org/abs/2505.09439), 2025.

665

666 S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ra-
 667 mani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio
 668 understanding and reasoning benchmark. In *The Thirteenth International Conference on Learn-
 669 ing Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

670

671 Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai
 672 Wang, and Shiguo Lian. DAST: difficulty-adaptive slow-thinking for large reasoning models.
CoRR, [abs/2503.04472](https://arxiv.org/abs/2503.04472), 2025.

673

674 Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and
 675 Chao Zhang. SALMONN: towards generic hearing abilities for large language models. In *The
 676 Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May
 7-11, 2024*, 2024.

677

678 Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer,
 679 Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal under-
 680 standing across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

681

682 Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

683

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

686

687 Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-
 688 reasoner: Improving reasoning capability in large audio language models. *CoRR*, [abs/2503.02318](https://arxiv.org/abs/2503.02318),
 689 2025.

690

691 Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang
 692 Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical
 693 report. *CoRR*, [abs/2503.20215](https://arxiv.org/abs/2503.20215), 2025.

694

695 Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. Text-to-audio grounding: Building corre-
 696 spondence between captions and sound events. In *IEEE International Conference on Acoustics,
 697 Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE,
 698 2021.

699

700 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang
 701 Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng
 702 Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang,
 703 Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu,
 704 Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men,
 705 Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren,

702 Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yingger Zhang,
 703 Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu.
 704 Qwen3 technical report. *CoRR*, abs/2505.09388, 2025.
 705

706 Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. AVQA:
 707 A dataset for audio-visual question answering on videos. In *MM '22: The 30th ACM International
 708 Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. ACM, 2022.

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

714 Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang,
 715 Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable
 716 lightweight infrastructure for fine-tuning. In *AAAI-25, Sponsored by the Association for the
 717 Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*. AAAI
 718 Press, 2025.
 719

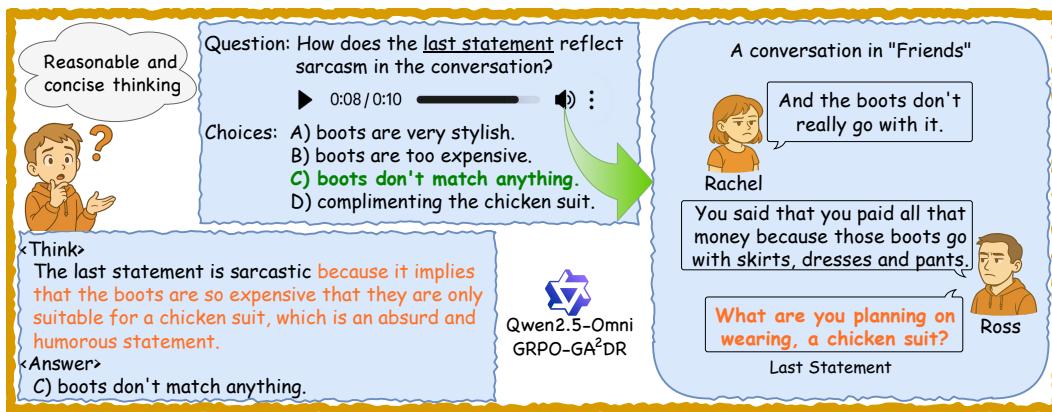
720

721 A APPENDIX

722

723 A.1 TASK DEFINITION

724



735 Figure 4: An audio QA example from “Friends”. The top-left shows the question and options
 736 (green indicates the correct one), the right side presents the audio dialogue, and the bottom-left
 737 shows the output of our proposed method on Qwen2.5-Omni-7B.
 738

743 A.2 SETUP

744

745 **Models.** For the base models, we primarily adopt Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-
 746 7B, which are among the most widely used open-source LALMs. These two models are also fre-
 747 quently used in prior work, facilitating fair and direct comparisons. For the training framework, we
 748 leverage the ms-swift (Zhao et al., 2025) and perform training on three A100-40G GPUs, where two
 749 GPUs are used for model training and one GPU is reserved for vLLM-based inference (Kwon et al.,
 750 2023). More training details and hyper-parameters can be found in the Appendix A.8.
 751

752 **Datasets.** For the training datasets, we use two in total: FS and AVQA (Yang et al., 2022). The
 753 latter is a subset of the former, while FS is constructed by augmenting AVQA with four additional
 754 datasets. For AVQA, we only keep the audio-text pairs and replace the word “video” in the ques-
 755 tions 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.

756 **GRPO.** GRPO has been extensively applied in both LLMs and MLLMs, achieving notable
 757 progress, and our implementation largely follows prior studies (DeepSeek-AI et al., 2025; Li et al.,
 758 2025). Compared with other RL methods, the key feature of GRPO is that it evaluates the policy
 759 model’s advantage using the average reward of in-group sampled outputs. Given an input question, a
 760 set of sampled responses for that question, and their rewards from the reward function, the advantage
 761 is calculated as follows:

$$762 \quad A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (6)$$

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

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

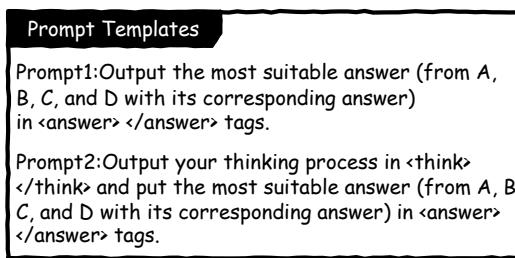
$$770 \quad \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|q)] \\ 771 \\ 772 \quad \frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathcal{D}_{KL}(\pi_{\theta} || \pi_{\text{ref}}) \right). \quad (7)$$

$$776 \quad \mathbb{D}_{\text{KL}}(\pi_{\theta} || \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i | q)}{\pi_{\theta}(o_i | q)} - \log \frac{\pi_{\text{ref}}(o_i | q)}{\pi_{\theta}(o_i | q)} - 1. \quad (8)$$

779 Here, ϵ and β are hyper-parameters.

781 **Cold-start GRPO.** A common way to learn the reasoning structure of advanced models and fur-
 782 ther optimize performance is to distill their outputs and then apply SFT to quickly teach the model
 783 such paradigms. After this, GRPO is performed, and this approach is referred to as Cold-Start. To
 784 explore the effectiveness of this method, we also conduct experiments with SFT Cold-Start followed
 785 by GRPO. For the Cold-Start dataset, we first sample a subset from the FS training set, distill it with
 786 Gemini2.5Pro, and then use Qwen3-235B-A22B³ to retain only samples that are both correct and
 787 consistent in reasoning and answers. From these, we select 200 per task to form a dataset of 1,000
 788 samples. This dataset is then used to perform Cold-Start SFT on Qwen2-Audio-7B-Instruct, and the
 789 model obtained after 2 epochs of SFT serves as the starting checkpoint for GRPO.

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



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

807
 808
 809 ³<https://huggingface.co/Qwen/Qwen3-235B-A22B>

810
 811 **Truncation Reward.** In addition to our designed rewards, we also consider the most straightforward
 812 way to control reasoning length as one of the baseline methods. This strategy sets a length
 813 threshold L_T , where rollout samples that are correct and have length less than or equal to L_T re-
 814 ceive a positive reward, while those exceeding it receive a penalty, even if the answer is correct. The
 815 calculation is as follows:

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

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

820 A.3 MODEL-PERSPECTIVE DIFFICULTY ON MMAU

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

825 Therefore, in this subsection, we introduce model-perspective difficulty labels. Since each model
 826 has different capabilities, we aim to reflect an average level across models. To this end, we adopt
 827 four different models: Qwen2-Audio-7B-Instruct, Qwen2.5-Omni-7B, Kimi-Audio-7B-Instruct⁴,
 828 and Gemini2.5-Pro-0506. These models perform inference on MMAU-test-mini under the same
 829 random seed, and difficulty labels from the model’s perspective are assigned based on the number of
 830 models answering each question correctly. The detailed distribution of model-perspective difficulty
 831 labels is shown in Table 9.

832 Table 9: Data distribution of difficulty from human (Orig) and model perspectives, including counts
 833 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

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

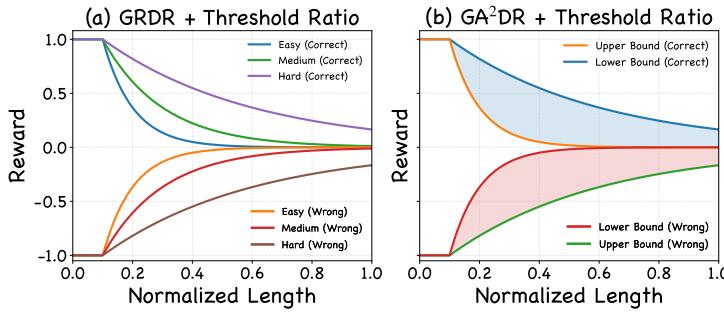
849 Considering that when the negative exponential function approaches 1 its exponent tends toward 0,
 850 this can represent a form of implicit reasoning, though it may reduce readability. To further analyze
 851 this issue, we introduce a parameter l_{min} , which serves as the minimum threshold ratio.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

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

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

859 Table 10, Table 11, and Table 12 report the effects of adding a threshold ratio to our two proposed
 860 rewards on three benchmarks. Models without special notation correspond to the case of $l_{min} =$
 861 0, which means that no threshold ratio is applied. For the cases with $l_{min} = 0.1$, this value is
 862 determined based on the average reasoning length of direct GRPO models.

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

Figure 6: Curves of GRDR and GA²DR with normalized length and the threshold ratio l_{min} .Table 10: The performance of our proposed GRDR, GA²DR, 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 [†]	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	71.47	71.85	66.66	57.75	92.78	80.58	59.81	60.34	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	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 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	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

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

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

924 In addition, when focusing on GRDR and GA²DR, we can observe that the latter achieves better
 925 performance on two benchmarks. This is because GA²DR defines difficulty levels at the batch
 926 level, which aligns with the batch-wise optimization process during model backpropagation, making
 927 this approach more effective. Moreover, its attention mechanism jointly considers both the textual
 928 question and the characteristics of the audio itself.

929 Furthermore, previous studies (Peng et al., 2025) have shown that models trained with reinforce-
 930 ment learning tend to exhibit improvements in pass@1 but declines in pass@k, largely due to the
 931 instability of rollout samples. Our GA²DR method effectively mitigates this issue, since all rollout
 932 samples within a group share the same question and audio input, thereby avoiding inconsistencies
 933 caused by rollout variance.

934 In summary, combining GA²DR with a length threshold provides the most balanced and effective
 935 optimization strategy. The specific threshold value, however, should be determined based on the
 936 characteristics of each individual task.

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

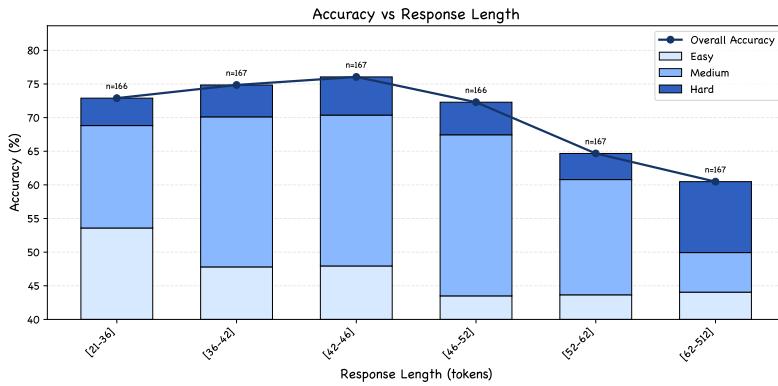
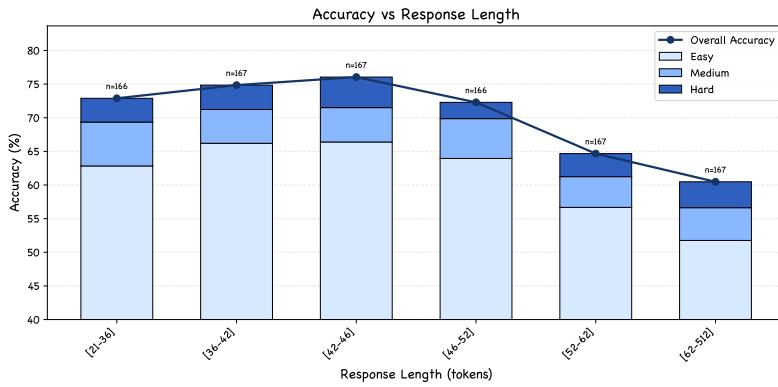
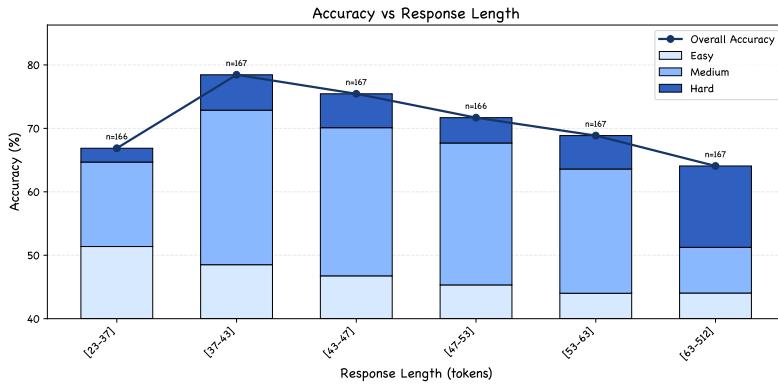
940 In this section, we provide the raw reasoning-length data for all models used in our main experiments
 941 on MMAU-Test-Mini, without any post-processing. All lengths are measured using the Qwen2-
 942 Audio tokenizer. Note that Qwen2-Audio and Qwen2.5-Omni share the same tokenizer, so using
 943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

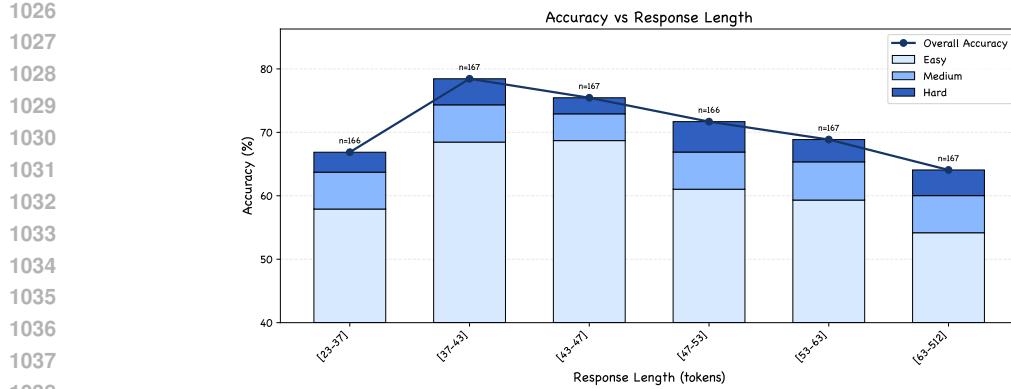
948 Table 13: The output length statistics of models on MMAU-test-mini with different base models
 949 and reward settings, reported under both the human-perspective difficulty annotations and the as-
 950 signed model-perspective difficulty annotations. Here, \dagger denotes the model-perspective difficulty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

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

972 A.6 REASONING LENGTH VS. ACCURACY
973

974 In this section, we analyze the relationship between output length and accuracy for the GRDR and
975 GA²DR methods on MMAU-Test-Mini. From Figure 7 to Figure 10, samples are grouped into
976 length intervals of similar size, and correctly answered samples within each interval are further
977 divided by difficulty level, indicated by increasingly darker shades of blue.

991 Figure 7: The trend of length and accuracy for the GRDR with human-perspective difficulty.
9921008 Figure 8: The trend of length and accuracy for the GRDR with model-perspective difficulty.
10091024 Figure 9: The trend of length and accuracy for the GA²DR with human-perspective difficulty.
1025

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

A.7 TRAINING CURVES

In this section, we present the training curves of GRDR and GA^2DR 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.

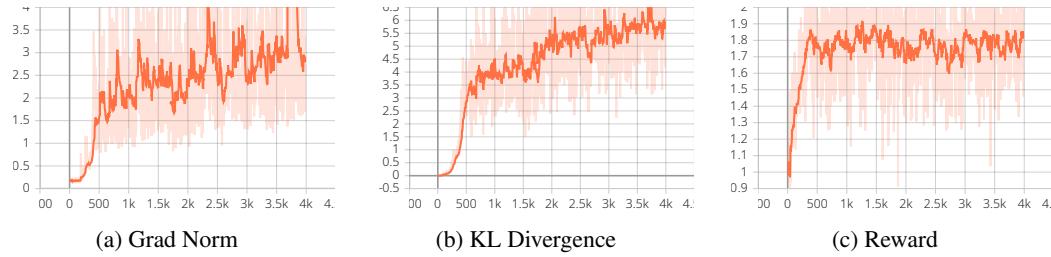
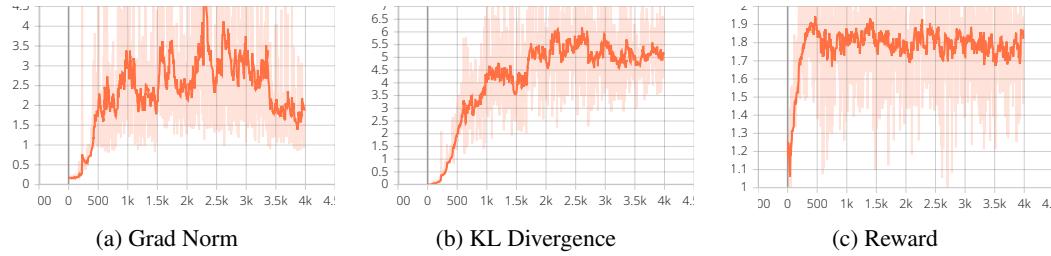


Figure 11: Training curves of gradient norm, KL divergence, and reward on GRDR.

Figure 12: Training curves of gradient norm, KL divergence, and reward on GA^2DR .

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.

1080 Table 14: Hyper-parameters for SFT and GRPO, including different settings under LoRA and Full.
1081

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 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

1105 **A.9 CASE STUDY: QUALITATIVE ANALYSIS OF OUTPUT PARADIGMS ACROSS MODELS**
1106

1107 In this section, we analyze output paradigms across models using 20 MMAU-test-mini questions on
1108 Dissonant Emotion Interpretation, focusing on sarcasm cause detection. The corresponding audios
1109 often involve multiple speakers and complex environments, making the task challenging and well-
1110 suited for qualitative analysis of different model outputs. In this section, in addition to the main
1111 models discussed above, we also include a Cold-Start GRPO model based on Qwen2-Audio-7B-
1112 Instruct, with detailed settings provided in the Appendix A.2.

1113 The detailed results are shown in Table 15, which include ACC and reasoning length on the sub-
1114 set mentioned above. In the second part, “Based on Qwen2-Audio-7B-Instruct,” we compare our
1115 three GRPO models with Audio-Reasoner. GRPO and GRPO with GRDR match Audio-Reasoner
1116 in performance while producing much shorter outputs, whereas the GRPO model after Cold-Start
1117 produces reasoning lengths close to Audio-Reasoner but performs worst. We believe this is mainly
1118 because the base model is weaker and can only imitate the surface structure of advanced model
1119 paradigms, showing that learning paradigms through Cold-Start is not effective in all cases and can
1120 even yield negative effects. In the “Based on Qwen2.5-Omni-7B” part of the table, performance
1121 differences are minor, mainly because the Qwen2.5-Omni base model already has some reason-
1122 ing ability and shows good performance on complex questions. However, with our two proposed
1123 rewards, the models maintain their original performance while significantly shortening reasoning
1124 length, greatly improving reasoning efficiency.

1125 In addition to analyzing performance and reasoning length above, we also compared the outputs
1126 of different models on these 20 questions in detail. From the results, we found that models with
1127 better performance usually follow a complete reasoning process, which includes first grounding the
1128 audio and providing a corresponding caption—equivalent to identifying the known conditions in the
1129 problem—then reasoning based on these conditions, analyzing each option step by step, and finally
1130 giving the answer. This approach helps LALMs make better use of known information together with
1131 their broad pretrained knowledge to perform reasonable reasoning while also improving readability.
1132 However, weaker models, even when using this structure, often make mistakes in the first step
1133 of extracting known information. These errors propagate through the reasoning process, leading
1134 to wrong answers and lower performance. For this part, we also provide an example of different
1135 models’ outputs in Appendix A.10 as reference.

1134 Table 15: Performance and reasoning length of different models on 20 sarcasm cause detection
 1135 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 ² DR	100	56.9
+ GA ² DR ($l_{min} = 0.1$)	95	94.2

1154 Overall, we believe that a good reasoning paradigm should consist of three parts: extracting known
 1155 conditions, reasoning and analysis, and giving the final answer. Among these steps, the earlier ones
 1156 have a stronger impact on later results, and once errors occur, they are difficult to correct in the
 1157 following steps. Therefore, we suggest that for weaker models, it is better to first fine-tune them on
 1158 basic tasks related to extracting known conditions (e.g., grounding and caption) before learning rea-
 1159 soning paradigms. For stronger models that still produce reasoning with poor readability, Cold-Start
 1160 can be used to learn the paradigm first, followed by RL training. For models that already generate
 1161 reasoning with good structure, direct RL with our two proposed rewards can further improve both
 1162 performance and efficiency.

A.10 EXAMPLE ANALYSIS

1165 In this section, we present output examples from six different models on a sarcasm cause detection
 1166 task from Figure 13 to Figure 15. This question is labeled as Medium in the original MMAU-test-
 1167 mini but is relabeled as Hard under our new scheme. The models include an advanced proprietary
 1168 model, Gemini2.5-Pro, the prior work Audio-Reasoner, and the four main models used in our ex-
 1169 periments. In the figure, the green text corresponds to the correct option for this example, the blue parts
 1170 in each model’s reasoning indicate logical and necessary steps or key points for solving the question,
 1171 and the red parts indicate repetition or redundancy, representing the parts that need optimization.

1188
1189
1190
1191
1192
1193

Example Analysis

1194
1195
1196
1197
1198
1199



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



⋮

1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

<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.



Gemini2.5-Pro

<Answer>

C) boots don't match anything.

<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).



Audio Reasoner

<Response>

The last statement ... implying that ... The answer is C.

<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.

<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.

<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.



Qwen2-Audio GRPO

<Answer>

C) boots don't match anything.

<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.



Qwen2-Audio Cold-Start GRPO

<Answer>

C) boots don't match anything.

<Think>

The last statement is sarcastic because it implies that the boots are so expensive that they are



Qwen2.5-Omni GRPO

<Answer>

C) boots don't match anything.



Qwen2.5-Omni GRPO-GA²DR

1242
1243
1244
1245
1246
1247

Example Analysis

1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295

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.

0:08 / 0:10



<Think>

The audio clip presents a conversation. Speaker 1: ... Speaker 2: ... The question asks why the last line ... can ... Sarcasm 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.



Gemini2.5-Pro

<Answer>

D) Not genuinely excited about assembling.

<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.



Audio Reasoner

<Response>

The speaker's tone ... The most ... The answer is C.

<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.



Qwen2-Audio GRPO

<Answer>

C) They are not coming over.

<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.



Qwen2-Audio Cold-Start GRPO

<Answer>

B) They don't know Joey and Chandler.

<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).



Qwen2.5-Omni GRPO

<Answer>

D) Not genuinely excited about assembling.

<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.



Qwen2.5-Omni GRPO-GA^2DR

<Answer>

D) Not genuinely excited about assembling.

1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349

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 🔍 ⏰

<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.

<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.



<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.



<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.



<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.



<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.