

# 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 EDGE-GRPO: ENTROPY-DRIVEN GRPO WITH GUIDED ERROR CORRECTION FOR ADVANTAGE DI- VERSITY

Anonymous authors

Paper under double-blind review

## ABSTRACT

Large Language Models (LLMs) have made remarkable progress in enhancing step-by-step reasoning through reinforcement learning. However, the Group Relative Policy Optimization (GRPO) algorithm, which relies on sparse reward rules, often encounters the issue of identical rewards within groups, leading to the advantage collapse problem. Existing works typically address this challenge from two perspectives: enforcing model reflection to enhance response diversity, and introducing internal feedback to augment the training signal (advantage). In this work, we begin by analyzing the limitations of model reflection and investigating the policy entropy of responses at the fine-grained sample level. Based on our experimental findings, we propose the EDGE-GRPO algorithm, which adopts **Entropy-Driven Advantage** and **Guided Error Correction** to effectively mitigate the problem of advantage collapse. Extensive experiments on different models across multiple main reasoning benchmarks demonstrate the effectiveness and superiority of our approach. The code and weights will be released upon acceptance to facilitate further research in this field.

## 1 INTRODUCTION

Recent advancements in large reasoning models, such as OpenAI-01 Jaech et al. (2024) and Kimi-K1.5 Team et al. (2025), have shown impressive progress in complex tasks involving mathematics and coding. Among them, the Group Relative Policy Optimization (GRPO) algorithm Shao et al. (2024) has attracted considerable attention from researchers. By discarding the value function used in the PPO algorithm Schulman et al. (2017) and instead computing rewards and relative advantages across sampled responses within each group, it significantly reduces resource consumption during training while improving reasoning performance.

For the computation of rewards in the GRPO algorithm, some studies adopt a Process Reward Model (PRM) to provide more fine-grained feedback Cui et al. (2025); Wang et al. (2025b). However, it introduces substantial computational overhead. As a result, other works abandon the reward model in favor of using rule-based reward functions Zhou et al. (2025); Zhang et al. (2025b). However, this often leads to sparse rewards, where all responses within a group receive identical rewards. Consequently, the calculated advantages for each response become zero, cease to provide

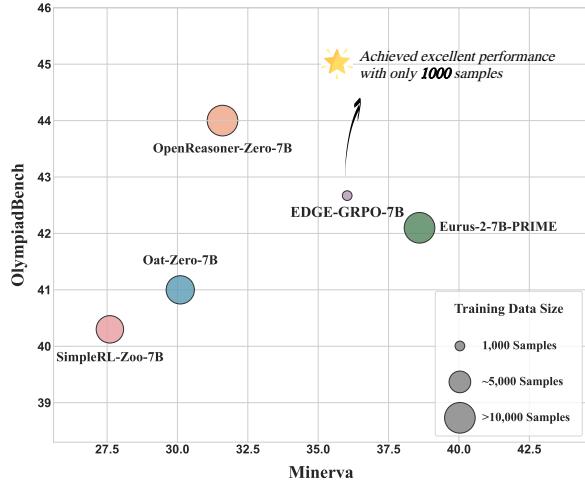


Figure 1: Performance comparison with other open-source models on Olympiad and Minerva. Our method achieves competitive and excellent performance with only 1K training samples. These models are all post-trained based on Qwen2.5-Math-7B.

054 effective policy gradient, and lead to advantage collapse during training. This phenomenon severely  
 055 limits the efficiency of the sample.  
 056

057 Recent studies have primarily sought to alleviate this problem from two perspectives. At the re-  
 058 sponse level, efforts focus on increasing the diversity of responses to prevent identical rewards across  
 059 all responses, such as enforcing model reflection on incorrect answers to reduce the occurrence of  
 060 uniformly incorrect outputs within a group Wang et al. (2025a); Wan et al. (2025). However, the  
 061 extent to which reflection contributes to performance improvement remains inconclusive. At the sig-  
 062 nal level, internal feedback is introduced to augment the advantage, such as incorporating response-  
 063 related semantic entropy or policy entropy into the advantage calculation Chen et al. (2025); Cheng  
 064 et al. (2025). However, most studies either pursue low entropy to improve accuracy or encourage  
 065 high entropy to maintain exploration, lacking fine-grained modeling of the relationship between  
 066 responses and their policy entropy.

067 In this paper, we begin by analyzing the limitations of model reflection. Quantitative experiments  
 068 show that responses containing self-reflection are often associated with significantly lower accuracy.  
 069 Although forced reflection can help the model correct a subset of answers, its overall effectiveness  
 070 remains limited. Additionally, we observe a misconception in the model’s estimation of policy  
 071 entropy at the fine-grained sample level: incorrect responses do not necessarily indicate uncertainty,  
 072 some of them exhibit notably lower entropy. Conversely, the model is not always confident in its  
 073 correct responses, some of which display relatively high entropy.

074 To address these issues, we propose a simple and effective EDGE-GRPO (Entropy-Driven GRPO  
 075 with Guided Error Correction) algorithm. At the response level, we introduce Guided Error Correc-  
 076 tion (GEC) to enhance response diversity, providing more effective guidance even when the model  
 077 encounters questions beyond its current capacity. At the signal level, we compute an Entropy-Driven  
 078 Advantage (EDA) that assigns higher advantages to correct responses with low entropy and lower  
 079 advantages to incorrect responses with low entropy, thereby increasing the diversity of the advan-  
 080 tage signal. These improvements significantly mitigate the problem of advantage collapse. Across  
 081 multiple reasoning benchmarks, our method achieves substantial performance gains compared to  
 082 the vanilla GRPO. As shown in Figure 1, our approach reaches comparable performance to other  
 083 open-source models using only 1K training samples.

084 Our contributions can be summarized as follows:

- 085 • We analyze the key challenges faced by preliminary attempts. Specifically, at the response  
 086 level, prompting the model to reflect on incorrect responses has limited effectiveness. At the  
 087 signal level, fine-grained sample-level policy entropy is needed to guide the augmentation  
 088 of the advantage.
- 089 • We propose the EDGE-GRPO algorithm. At the response level, we introduce Guided Error  
 090 Correction (GEC) to overcome the limitations of the model capacity and improve response  
 091 diversity. At the signal level, we compute an Entropy-Driven Advantage (EDA) to increase  
 092 the diversity of the advantage signal, significantly alleviating the problem of advantage  
 093 collapse.
- 094 • Extensive experiments on multiple main reasoning benchmarks show that our method  
 095 achieves a significant performance improvement across different model families and sizes,  
 096 thus validating its effectiveness and superiority.

## 099 2 RELATED WORK

100 **Advantage Collapse.** Advantage collapse is a critical limitation of the GRPO algorithm, as it  
 101 severely impairs effective gradient updates. Prior approaches typically mitigate this issue through  
 102 data filtering Yu et al. (2025); Meng et al. (2025), by discarding samples in which all responses  
 103 within a group are either entirely correct or incorrect. However, this greatly limits sample effi-  
 104 ciency, as challenging samples can be beneficial for improving model performance. In addition,  
 105 some works Wang et al. (2025a) attempt to enhance response diversity by enforcing model reflec-  
 106 tion, while others Chen et al. (2025); Cheng et al. (2025) introduce internal feedback to strengthen  
 107 the training signal.

108 **Think More or Less.** There are differing views on whether model reflection truly benefits model  
 109 performance. Several works Muennighoff et al. (2025); Tian et al. (2025) proposed adding "wait"  
 110 to chain-of-thought reasoning to encourage the model to engage in reflection, which can improve  
 111 performance. VL-Rethinker Wang et al. (2025a) incorporates forced reflection during the training  
 112 process to enhance the slow-thinking capability of the model. Meanwhile, other researchers argue  
 113 that suppressing the tokens that trigger reflection Liu et al. (2025a), encouraging the model to generate  
 114 shorter responses Su et al. (2025); Fatemi et al. (2025), can reduce redundant reasoning without  
 115 compromising the model's accuracy.

116 **RL from Internal Feedback.** Recent studies introduce internal feedback such as entropy to  
 117 strengthen the training signal. Some studies Gao et al. (2025); Zhang et al. (2025a) argue that  
 118 correct responses generated by models typically exhibit lower entropy than incorrect ones, so unsupervised  
 119 entropy minimization methods can also enhance performance. SEED-GRPO Chen et al.  
 120 (2025) introduces semantic entropy to quantify semantic diversity among generated responses and  
 121 dynamically adjusts the magnitude of policy updates based on this measure. Other works Cheng  
 122 et al. (2025) suggest that high entropy encourages exploratory reasoning, therefore incorporating  
 123 policy entropy into the advantage term of the GRPO algorithm to promote exploration. However,  
 124 most of these methods lack fine-grained modeling of the relationship between response correctness  
 125 and their policy entropy.

### 127 3 INVESTIGATION OF ADVANTAGE COLLAPSE IN GRPO

128 We begin with a brief introduction to the Group Relative Policy Optimization (GRPO) algorithm  
 129 Shao et al. (2024). For each input question  $q$ , it generates a set of responses  $\{O_1, O_2, \dots, O_G\}$   
 130 using the policy model and computes a corresponding set of rewards  $\{r_1, r_2, \dots, r_G\}$  for these  
 131 responses. The rewards are then normalized to calculate the advantages. The model is optimized by  
 132 maximizing the following objective function:

$$134 \quad J_{GRPO}(\theta) = E_{[q, \{o_i\}]} \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{i=1}^{|o_i|} \left\{ \min \left[ \frac{\pi_\theta}{\pi_{\theta_{old}}} A_i, \text{clip} \left( \frac{\pi_\theta}{\pi_{\theta_{old}}}, 1 - \epsilon, 1 + \epsilon \right) A_i \right] - \beta D_{KL} \right\} \quad (1)$$

135 where  $\pi_\theta$  and  $\pi_{\theta_{old}}$  are the current and old policy, and  $A_i$  is the advantages defined as:

$$136 \quad A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (2)$$

137 The diversity of advantages is crucial for effective model updates, as it directly determines the  
 138 training signal used in policy gradient optimization. Due to the difficulty of assigning rewards to  
 139 intermediate reasoning steps, most existing reward rules are sparse, with the response reward largely  
 140 determined by the correctness of the final answer. As a result, when all responses within a group  
 141 are either correct or incorrect, they receive identical rewards, leading to zero advantage across the  
 142 group. This lack of distinction between responses impairs gradient updates, a phenomenon known  
 143 as the advantage collapse problem.

144 Advantage collapse results in low sample efficiency. However, samples that are more challenging  
 145 for the model often play an important role in improving its performance. Therefore, addressing the  
 146 advantage collapse problem remains a critical challenge.

147 Existing approaches commonly aim to address this issue from two key perspectives: at the response  
 148 level, by promoting model reflection to enhance the diversity of generated responses. At the signal  
 149 level, by incorporating internal feedback mechanisms to enrich the training signal. In this work, we  
 150 first conduct a preliminary investigation along both dimensions to explore their potential in mitigat-  
 151 ing the advantage collapse problem.

#### 152 3.1 RESPONSE-LEVEL: LIMITATIONS OF REFLECTION

153 We begin by conducting a series of quantitative experiments to analyze the phenomenon of model  
 154 reflection. We select models of two different parameter scales, including base models as well as  
 155 those post-trained by supervised fine-tuning or reinforcement learning. To determine whether a

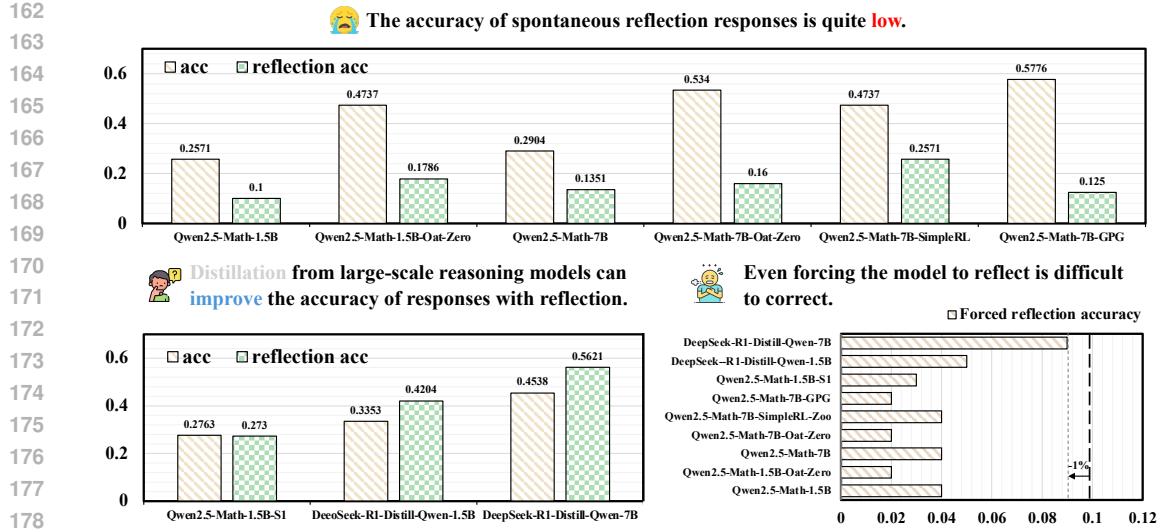


Figure 2: The reflection performance of different models. **Upper:** For most models, the accuracy of responses that involve self-reflection is significantly lower than the overall accuracy. **Left:** Fine-tuning with high-quality data that includes reflection processes helps improve the accuracy of model reflection. **Right:** Even when the model is forced to reflect on incorrect responses, the improvement in accuracy remains limited, these results are averaged over four types of reflection prompts.

model’s response exhibits self-reflection, we follow previous work Liu et al. (2025b) by extracting reflection-related keywords from the responses. If the response contains reflection keywords such as `check again`, it is considered to exhibit self-reflection. The specific set of reflection keywords is provided in the Appendix.

Initially, we observe that the majority of spontaneously generated reflections by models tend to exhibit low accuracy. As shown in the upper part of Figure 2, for both base models and those post-trained by reinforcement learning, the accuracy of responses containing reflection is significantly lower than the overall accuracy of the model. This result clearly indicates that spontaneous reflection during reasoning is often ineffective and may even lead to a higher rate of incorrect responses.

However, unlike other models, two models distilled from DeepSeek-R1 Guo et al. (2025) have a more frequent self-reflection behavior, and their reflection is accompanied by higher accuracy. To verify whether this phenomenon is caused by long chain-of-thought training data from knowledge distillation, we train Qwen2.5-Math-1.5B-S1 on the S1K dataset Muennighoff et al. (2025), which contains only 1K high quality long chain-of-thought samples, some of which include reflection-related content. It can be observed that after training on the S1K data, the model’s reflection accuracy significantly improves and becomes comparable to its overall accuracy.

Subsequently, we also investigate the effect of forcing different models to reflect on their incorrect answers. Specifically, we first have each model respond to every question in the test set, then we retain only the samples with incorrect answers. A reflection prompt is appended to each incorrect response to initiate reflection, after which the model is prompted to continue answering. We designed four distinct reflection prompts: `Wait!`, `Hmm`, `Let’s check it again!`, and `Something is wrong here`. These prompts include two anthropomorphic expressions and two objective declarative phrases.

As shown in the lower right corner of Figure 2, the accuracy of forced reflection on incorrect responses remains below 5% for most models. Although the overall accuracy does not exceed 10%, the DeepSeek-R1-Distill series model still achieves relatively higher accuracy compared to other models due to being fine-tuned with external high-quality chain-of-thought data.

These results reveal a fundamental limitation in the reflective capabilities of most models. When model capacity is limited, relying solely on self-correction yields minimal improvement. Therefore, when confronted with challenging problems where the model persistently produces incorrect

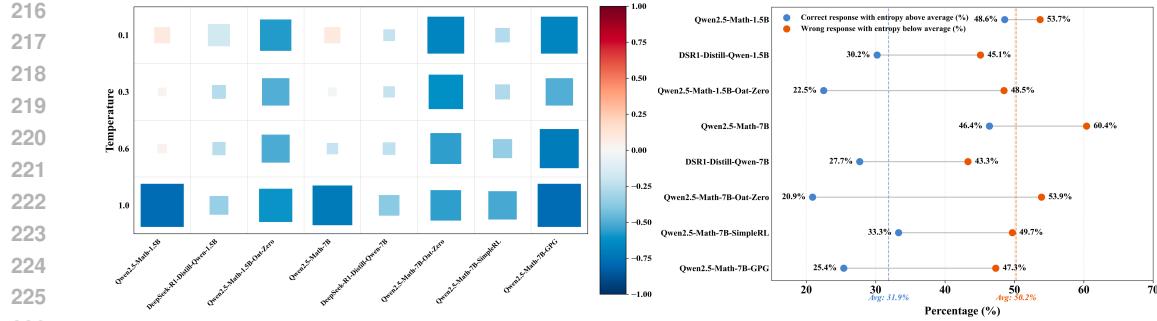


Figure 3: **Left:** The relative confidence of different models in correct responses under various temperature settings. The area of the blue squares serves as a proxy for the model’s relative confidence, with larger areas reflecting greater confidence in correct responses. **Right:** The proportion of correct responses with entropy higher than the average and incorrect responses with entropy lower than the average across different models. These results are evaluated under the setting of temperature=0.1. We provide more detailed experimental results and the policy entropy distribution of different models in the Appendix.

responses, incorporating external information for correction emerges as a more effective and reliable strategy at the response level.

### 3.2 SIGNAL-LEVEL: POLICY ENTROPY

We also investigate the policy entropy of different responses. For each generated chain-of-thought response, the policy entropy  $P$  is calculated as follows:

$$P = -\frac{1}{T} \sum_{t=1}^T \sum_{j=1}^V P_{t,j} \cdot \log P_{t,j}. \quad (3)$$

where  $T$  denotes the total number of tokens in the response,  $V$  is the vocabulary size, and  $P_{t,j}$  is defined as:

$$P_{t,j} = \pi_\theta(j \in V | q, o < t) = \text{Softmax} \left( \frac{\text{logits}_t}{T} \right). \quad (4)$$

Here,  $\pi_\theta$  represents the language model parameterized by  $\theta$ . We use the policy entropy  $P$  to measure the uncertainty of the model over the generated response.

We first divide all responses into two categories based on whether they are correct or incorrect, and calculate the Relative Confidence Metric (RCM) of each model in correct responses under different temperatures using the following formula:

$$\text{RCM} = \frac{\text{Entropy}_{\text{Correct}} - \text{Entropy}_{\text{Wrong}}}{\text{Average Entropy}}. \quad (5)$$

The visualization results are shown on the left side of Figure 3. Except for the two base models, Qwen2.5-1.5B and Qwen2.5-7B Yang et al. (2024), other post-trained models typically exhibit higher relative confidence, as the average entropy of their correct responses is indeed lower than that of incorrect responses, which aligns with assumptions made in previous studies Gao et al. (2025).

However, a fine-grained analysis at the individual sample level reveals that many models exhibit miscalibrated confidence in their responses: approximately half of the incorrect responses display entropy values lower than the average, while nearly one-third of the correct responses exhibit entropy higher than the average, as shown on the right side of Figure 3. We posit that such miscalibration undermines model performance. Ideally, the model should exhibit greater confidence in the correct responses while maintaining appropriate uncertainty about its incorrect answers. Consequently, training strategies should avoid indiscriminately promoting high or low entropy, and instead adopt a more fine-grained approach that aligns policy entropy with response correctness to better guide learning dynamics.

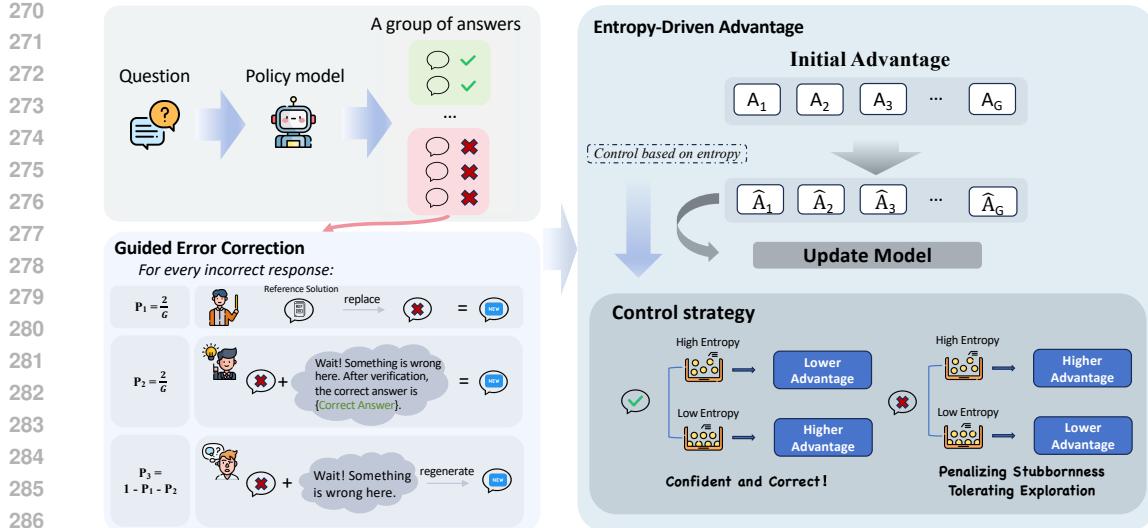


Figure 4: The overall framework of EDGE-GRPO. By introducing Guided Error Correction at the response level to enhance response diversity and Entropy-Driven Advantage at the signal level to increase advantage diversity, we mitigate the advantage collapse problem in the vanilla GRPO. Here,  $G$  represents the number of generated responses in a group.

## 4 EDGE-GRPO: ENTROPY-DRIVEN GRPO WITH GUIDED ERROR CORRECTION

Building on the above insights, we propose the EDGE-GRPO algorithm by introducing Guided Error Correction (GEC) to enhance response diversity and Entropy-Driven Advantage (EDA) to augment signal diversity, thus addressing the advantage collapse problem.

### 4.1 RESPONSE-LEVEL: GUIDED ERROR CORRECTION (GEC)

The experimental analysis in the previous section has shown that the model’s ability to correct errors through reflection is quite limited. This limitation leads the model to consistently generate entirely incorrect responses when faced with problems beyond its capabilities. However, response diversity fundamentally impacts reward diversity, which consequently directly affects advantage diversity. Therefore, introducing external solutions to ensure that each set of responses contains a certain proportion of positive and negative samples is crucial for mitigating the advantage collapse problem.

To fundamentally address this issue, we propose Guided Error Correction (GEC), a response-level intervention strategy designed to mitigate advantage collapse by enhancing response diversity. As illustrated in Figure 4, for incorrect responses, GEC performs one of the following three operations based on a predefined probability:

**Reference Solution Replacement:** The incorrect response is completely replaced with an external reference solution. This operation is performed with a probability of  $P_1 = \frac{2}{G}$ , where  $G$  is the total number of responses in the group. Compared to a smaller value ( $P_1 = \frac{1}{G}$ ), this value ensures a higher probability that at least one response will be replaced with the reference solution when all responses are incorrect. Compared to a higher probability, it also maintains response diversity and avoids having too many identical responses within the group.

**Direct Answer Injection:** Along with the reflection prompt, the correct answer is provided directly. This operation is also performed with a probability of  $P_2 = \frac{2}{G}$  to ensure both the number and the diversity of correct responses within the group.

**Prompt and Regenerate:** A simple reflection prompt is provided, and the model is asked to regenerate its answer based on it, giving the model a chance to self-correct. This operation is performed

324  
 325 Table 1: Pass@1 performance comparison across various mathematical evaluation benchmarks. The  
 326 results below are from 1 epoch of training on **DeepScaleR-Hard-1K**. The number of samples in each  
 327 benchmark is indicated in parentheses. The results are evaluated under the setting of temperature =  
 0.1. The best results are indicated by **boldface**.

Model	Method	Avg (1560)	AIME (30)	AMC (83)	Math (500)	Min (272)	Oly (675)
Qwen2.5-Math-1.5B	Base	25.71	6.67	37.35	34.60	12.13	24.00
	SFT	29.17	6.67	28.92	46.40	12.87	24.00
	Vanilla GRPO	40.26	10.00	<b>46.99</b>	65.00	20.59	30.37
	+ Force Reflection	41.55	<b>13.33</b>	31.33	70.00	22.06	30.81
	Dr.GRPO	40.13	13.33	43.37	67.20	18.75	29.48
	DAPO	33.91	6.67	31.33	54.80	16.54	26.96
Qwen2.5-Math-7B	<b>EDGE-GRPO</b>	<b>47.24</b>	10.00	44.58	<b>73.20</b>	<b>29.04</b>	<b>37.33</b>
	Base	29.04	10.00	37.35	53.40	10.66	18.22
	SFT	37.37	3.33	44.58	68.00	19.85	22.37
	Vanilla GRPO	47.69	<b>26.67</b>	<b>53.01</b>	74.20	25.74	37.19
	+ Force Reflection	40.26	13.33	40.96	67.60	19.85	29.33
	Dr.GRPO	48.78	23.33	56.63	75.20	27.21	38.07
Llama-3.2-3B-Instruct	DAPO	49.49	20.00	57.83	76.20	27.94	38.67
	<b>EDGE-GRPO</b>	<b>53.21</b>	16.67	<b>53.01</b>	<b>79.00</b>	<b>36.03</b>	<b>42.67</b>
	Base	19.81	6.67	14.46	36.20	12.13	12.00
	SFT	22.44	0.00	20.48	41.20	11.77	14.07
	Vanilla GRPO	22.44	<b>13.33</b>	15.66	42.80	15.07	11.56
	+ Force Reflection	22.89	6.67	<b>21.69</b>	41.8	13.97	13.33
Qwen2.5-Math-7B	Dr.GRPO	22.24	3.33	18.07	42.60	12.50	12.44
	DAPO	22.95	3.33	18.07	42.60	13.97	13.48
	<b>EDGE-GRPO</b>	<b>25.06</b>	3.33	20.48	<b>45.60</b>	<b>17.28</b>	<b>14.52</b>

345  
 346 with a probability of  $P_3 = 1 - P_1 - P_2$ , which ensures that most responses are still generated by  
 347 the model itself. Since the effect of self-reflection is relatively limited, only a small portion of the  
 348 responses can be corrected to the right answers. These few corrected responses are then used as  
 349 positive samples to guide the model update.

350 These three strategies ensure that each group of responses contains positive samples with correct  
 351 answers while still retaining negative samples generated by the model itself.

352 By introducing Guided Error Correction at the response level, we ensure that even when the model  
 353 encounters problems beyond its capabilities, the response set can still contain diverse answers. This  
 354 helps mitigate the issue of advantage collapse and provides effective training signals.

#### 356 4.2 SIGNAL-LEVEL: ENTROPY-DRIVEN ADVANTAGE (EDA)

357  
 358 Although Guided Error Correction enhances response diversity and prevents the advantages within  
 359 a group from collapsing to zero, it remains insufficient to address the issue of uniform advantages  
 360 among correct or incorrect responses. To enable finer-grained differentiation among different cor-  
 361 rect or incorrect responses, we introduce policy entropy as an internal feedback signal to enhance  
 362 advantage diversity.

363 The results in the previous section show that the model often misjudges the confidence of its re-  
 364 sponses, many incorrect responses exhibit low entropy, while many correct responses have high en-  
 365 tropy. We believe this misalignment negatively impacts model performance. Therefore, we propose  
 366 Entropy-Driven Advantage (EDA) to enhance the model’s ability to distinguish between different  
 367 responses.

368 For each response  $O_i$  during training, we calculate its policy entropy  $P_i$  using Equations 3 and 4,  
 369 and then scale it to ensure the values remain within a reasonable range.

$$370 \hat{P}_i = \frac{P_i}{\text{mean}(\{P_1, P_2, \dots, P_G\})}. \quad (6)$$

371 Next, we use the scaled entropy values to compute the entropy-driven advantage:

$$372 \hat{A}_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\}) \cdot \hat{P}_i}. \quad (7)$$

373  
 374 Compared to the initial advantage values, the entropy-driven advantage exhibits greater diversity. It  
 375 assigns higher advantages to responses that are both correct and confident, while imposing harsher

378 penalties on responses that are incorrect but overly confident. It ensures that, when the initial ad-  
 379 vantages are not all zero, different responses are assigned distinct final advantage values, thereby  
 380 enhancing the model’s ability to distinguish among responses and further mitigating the advantage  
 381 collapse problem.

382 It is worth emphasizing that the GEC and EDA modules enhance the diversity of advantage from  
 383 different levels, and they are complementary to some extent. Although the GEC module ensures  
 384 that the advantage of the response is not zero, it cannot achieve finer-grained differentiation between  
 385 correct or incorrect responses. In contrast, the EDA module relies on the presence of a certain  
 386 number of positive and negative samples within the group. When the initial intra-group advantages  
 387 are all zero, fine-grained scaling cannot take effect. This is precisely the reason for incorporating a  
 388 proportion of correct responses into the GEC module.

## 390 5 EXPERIMENTS

### 391 5.1 EXPERIMENTAL SETUP

394 **Train Datasets.** We use the DeepScaleR dataset Luo et al. (2025) for training. The original dataset  
 395 contains approximately 40K math problems. We retain only those samples that include a solution  
 396 and where the final answer is placed inside a `\boxed{}` in the solution. After this filtering process,  
 397 around 2K samples remain. We randomly select 1K samples as the standard training set, named  
 398 DeepScaleR-Random-1K. Meanwhile, to evaluate the effectiveness of our method on more chal-  
 399 lenging data, we use Qwen2.5-Math-7B to further filter the samples. Specifically, for each question,  
 400 the model generates eight responses, and we select the 1K questions with the lowest accuracy as  
 401 the hard training set, referred to as DeepScaleR-Hard-1K. In this dataset, approximately 80% of the  
 402 questions receive entirely incorrect responses across all generations.

403 **Evaluation Benchmark.** We select five challenging mathematical reasoning benchmarks to eval-  
 404 uate our method: AIME24, AMC, MATH500 Hendrycks et al. (2021), Minerva Lewkowycz et al.  
 405 (2022) and OlympiadBench He et al. (2024). These benchmarks collectively contain a total of 1,560  
 406 problems. All evaluation experiments in this paper are conducted on these benchmarks.

408 **Implementation Details.** We conduct experiments on 8 NVIDIA A100-40G GPUs. We remove  
 409 the KL divergence to eliminate constraints on the model. Previous studies Yu et al. (2025); Liu et al.  
 410 (2025b) have shown that it can lead to better training performance, as the distribution of the model  
 411 may differ significantly from the initial model during training. Other training configurations and  
 412 hyperparameter settings follow the default setup of the GRPO trainer under the TRL framework von  
 413 Werra et al. (2020). We train for one epoch on only 1K DeepScaleR samples on Qwen2.5-Math-  
 414 1.5B, Qwen2.5-Math-7B Yang et al. (2024) and Llama3.2-3B-Instruct Grattafiori et al. (2024).

415 During evaluation, we focus on the model’s pass@1 performance, meaning the model generates only  
 416 one response for each given question. To calculate the overall average accuracy, we avoid directly  
 417 averaging the accuracy across the five benchmarks due to their varying number of questions. Instead,  
 418 we calculate the average by dividing the total number of correct answers by the total number of  
 419 questions to reduce bias. More detailed experimental settings can be found in the Appendix.

### 421 5.2 MAIN RESULT

423 Table 1 presents the results of our method on various mathematical evaluation benchmarks. Al-  
 424 though our method is trained on only 1K samples for one epoch, it achieves significant performance  
 425 improvements across various scales of Qwen and Llama models. It is worth emphasizing that for  
 426 about 80% of the problems in DeepScaleR-Hard-1K, all eight responses generated by the base model  
 427 Qwen2.5-Math-7B are incorrect. This further validates the effectiveness of our method in challeng-  
 428 ing data, even when the difficulty exceeds the capability of the model. In addition, we also conduct  
 429 experiments on DeepScaleR-Random-1K to verify the generalization performance of our method,  
 430 with detailed results provided in the Appendix.

431 Since our method requires each question to have not only the final answer, but also a corresponding  
 432 reference solution, we established a Supervised Fine-Tuning (SFT) baseline using the same chain-

432 Table 2: The ablation study of EDGE-GRPO separately verifies the effectiveness of guided error  
 433 correction and entropy-driven advantage. The results are all from training on DeepScaleR-Hard-1K.  
 434

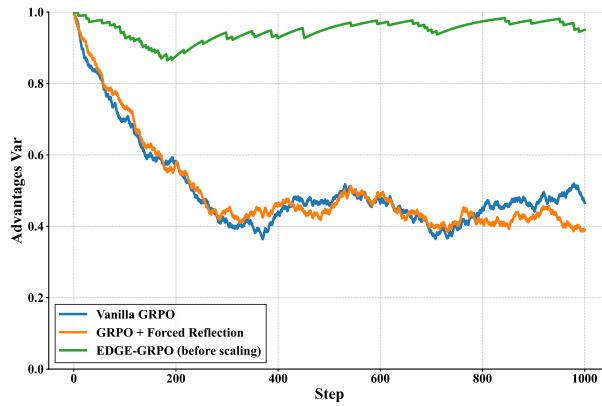
435 <b>Method</b>	436 <b>Avg (1560)</b>	437 <b>AIME (30)</b>	438 <b>AMC (83)</b>	439 <b>Math (500)</b>	440 <b>Min (272)</b>	441 <b>Oly (675)</b>
<b>EDGE-GRPO [Qwen2.5-Math-1.5B]</b>	47.24	10.00	44.58	73.20	29.04	37.33
- Reference Solution	42.44	10.00	34.94	69.60	23.53	32.30
- Guided Error Correction	40.06	13.33	40.96	64.00	21.32	30.96
- Entropy-Driven Advantage	40.64	10.00	43.37	67.80	19.12	30.22
<b>EDGE-GRPO [Qwen2.5-Math-7B]</b>	53.21	16.67	53.01	79.00	36.03	42.67
- Reference Solution	45.19	26.67	50.60	68.00	28.68	35.11
- Guided Error Correction	46.80	20.00	50.60	75.80	23.16	35.56
- Entropy-Driven Advantage	47.44	3.33	43.37	74.40	29.04	37.33

444 of-thought data for comparison. As shown in the SFT row of Table 1, our approach significantly  
 445 outperforms supervised fine-tuning, even when trained on the exact same data. And our method also  
 446 exhibits a clear advantage when evaluated against established algorithmic variants. Specifically,  
 447 our method consistently outperforms algorithmic variants such as Dr.GRPO Liu et al. (2025b) and  
 448 DAPO Yu et al. (2025), as well as improvements to the vanilla GRPO algorithm that force the model  
 449 to reflect on incorrect responses.

450 In addition, we conduct an ablation study on EDGE-GRPO, as shown in Table 2. First, removing  
 451 external reference solutions leads to a decline in model performance, highlighting the importance  
 452 of integrating external information when dealing with problems beyond the model’s capability. In  
 453 addition, removing either the GEC or EDA component results in a significant performance gap  
 454 compared to EDGE-GRPO, which not only validates the effectiveness of each component but also  
 455 underscores the importance of enhancing response diversity and providing more fine-grained dis-  
 456 tinctions in training signals for improving model performance.

457 We also visualize changes in advantage variance during training for different methods, as shown in Figure 5. During  
 458 training, compared with the vanilla GRPO algorithm and the variant with  
 459 enforced reflection, our method maintains a higher level of intra-group advantage variance solely through response-  
 460 level improvements (GEC). After advantage scaling, advantage diversity will be  
 461 further enhanced. It also demonstrates its significant mitigation of the advantage collapse problem.

462 Moreover, even with significantly less  
 463 training data, our model achieves comparable  
 464 performance to other main open  
 465 source models Zeng et al. (2025); Hu  
 466 et al. (2025); Cui et al. (2025); Liu et al.  
 467 (2025b), as illustrated in Figure 1. This further demonstrates the superiority and utility of our  
 468 method. More detailed experimental results can be found in the Appendix.



469 Figure 5: The changes in intra-group advantage variance  
 470 during training for different methods. Our method maintains a relatively high level without significant decline.

## 471 6 CONCLUSION

472 This work proposes a simple and effective EDGE-GRPO algorithm that mitigates the advantage  
 473 collapse problem of the vanilla GRPO algorithm on two levels. At the response level, the Guided  
 474 Error Correction (GEC) method is introduced to overcome the limitations of the inherent capabilities  
 475 of the model and improve response diversity. At the signal level, the Entropy-Driven Advantage  
 476 (EDA) computation enables the model to differentiate responses more finely during training, thereby  
 477 improving the diversity of advantages. Our method significantly alleviates the advantage collapse  
 478 problem and achieves notable performance improvements using only 1K samples across different  
 479 base models, demonstrating its effectiveness and superiority.

486 REFERENCES  
487

488 Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced  
489 grp for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*, 2025.

490 Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and  
491 Furu Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*,  
492 2025.

493 Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu  
494 Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint  
495 arXiv:2502.01456*, 2025.

496 Mehdi Fatemi, Banafsheh Rafiee, Mingjie Tang, and Kartik Talamadupula. Concise reasoning via  
497 reinforcement learning. *arXiv preprint arXiv:2504.05185*, 2025.

498 Zitian Gao, Lynx Chen, Joey Zhou, and Bryan Dai. One-shot entropy minimization. *arXiv preprint  
499 arXiv:2505.20282*, 2025.

500 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad  
501 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd  
502 of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

506 Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu,  
507 Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for  
508 promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint  
509 arXiv:2402.14008*, 2024.

510 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song,  
511 and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv  
512 preprint arXiv:2103.03874*, 2021.

513 Jingcheng Hu, Yinmin Zhang, Qi Han, Dixin Jiang, Xiangyu Zhang, and Heung-Yeung Shum.  
514 Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base  
515 model. *arXiv preprint arXiv:2503.24290*, 2025.

516 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec  
517 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv  
518 preprint arXiv:2412.16720*, 2024.

519 Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay  
520 Ramasesh, Ambrose Sloane, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quan-  
521 titative reasoning problems with language models, 2022. URL <https://arxiv.org/abs/2206.14858>,  
522 1, 2022.

523 Kaiyuan Liu, Chen Shen, Zhanwei Zhang, Junjie Liu, Xiaosong Yuan, et al. Efficient reasoning  
524 through suppression of self-affirmation reflections in large reasoning models. *arXiv preprint  
525 arXiv:2506.12353*, 2025a.

526 Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee,  
527 and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint  
528 arXiv:2503.20783*, 2025b.

529 Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Tang, Manan Roongta, Colin Cai,  
530 Jeffrey Luo, Tianjun Zhang, Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing  
531 o1-preview with a 1.5b model by scaling rl, 2025. Notion Blog.

532 Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi,  
533 Wenhui Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with  
534 rule-based large-scale reinforcement learning. *CoRR*, 2025.

540 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke  
 541 Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time  
 542 scaling. *arXiv preprint arXiv:2501.19393*, 2025.

543

544 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy  
 545 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

546

547 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,  
 548 Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical  
 549 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

550

551 Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. Between underthinking and over-  
 552 thinking: An empirical study of reasoning length and correctness in llms. *arXiv preprint  
 553 arXiv:2505.00127*, 2025.

554

555 Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun  
 556 Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with  
 557 llms. *arXiv preprint arXiv:2501.12599*, 2025.

558

559 Xiaoyu Tian, Sitong Zhao, Haotian Wang, Shuaiting Chen, Yunjie Ji, Yiping Peng, Han Zhao, and  
 560 Xiangang Li. Think twice: Enhancing llm reasoning by scaling multi-round test-time thinking.  
 561 *arXiv preprint arXiv:2503.19855*, 2025.

562

563 Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan  
 564 Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement  
 565 learning. <https://github.com/huggingface/trl>, 2020.

566

567 Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong,  
 568 Yi Xin, Yifan Jiang, et al. Srpo: Enhancing multimodal llm reasoning via reflection-aware rein-  
 569 force learning. *arXiv preprint arXiv:2506.01713*, 2025.

570

571 Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. Vi-  
 572 rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning.  
 573 *arXiv preprint arXiv:2504.08837*, 2025a.

574

575 Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu,  
 576 Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for  
 577 multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025b.

578

579 An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jian-  
 580 hong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical  
 581 expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

582

583 Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian  
 584 Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system  
 585 at scale. *arXiv preprint arXiv:2503.14476*, 2025.

586

587 Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-  
 588 zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv  
 589 preprint arXiv:2503.18892*, 2025.

590

591 Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question  
 592 is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint  
 593 arXiv:2504.05812*, 2025a.

594

595 Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller  
 596 llms for video reasoning. *arXiv preprint arXiv:2504.09641*, 2025b.

597

598 Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-  
 599 zero’s” aha moment” in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*,  
 600 2025.

601

594  
595  
596  
597  
598  
A APPENDIX599  
600  
601  
602  
603  
A.1 REFLECTION KEYWORDS604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
We determine whether a reflection phenomenon has occurred based on the presence of reflection-related keywords in the response. The set of 15 keywords used in the experiment for identifying reflection is as follows: check again, recheck, double-check, rethink, think again, reevaluate, re-evaluate, re-examine, verify again, reevaluation, reexamine, reanalyze, reassess, reconsider, go over.617  
618  
619  
620  
621  
622  
623  
624  
A.2 ANALYSIS ON THE INTRODUCTION OF EXTERNAL SOLUTIONS625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
In the Guided Error Correction (GEC) section, we introduce external reference solutions to enhance response diversity. Since the reference solution trajectories are not sampled by the old policy, our method is not strictly on-policy. So we approximate the probability of these solutions under the old model policy  $\pi_{\theta_{old}}$  using the probability that the old model would generate the trajectories of these reference solutions. The specific probability calculation is shown in Algorithm 1.648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
10010  
10011  
10012  
10013  
10014  
10015  
10016  
10017  
10018  
10019  
10020  
10021  
10022  
10023  
10024  
10025  
10026  
10027  
10028  
10029  
10030  
10031  
10032  
10033  
10034  
10035  
10036  
10037  
10038  
10039  
10040  
10041  
10042  
10043  
10044  
10045  
10046  
10047  
10048  
10049  
10050  
10051  
10052  
10053  
10054  
10055  
10056  
10057  
10058  
10059  
10060  
10061  
10062  
10063  
10064  
10065  
10066  
10067  
10068  
10069  
10070  
10071  
10072  
10073  
10074  
10075  
10076  
10077  
10078  
10079  
10080  
10081  
10082  
10083  
10084  
10085  
10086  
10087  
10088  
10089  
10090  
10091  
10092  
10093  
10094  
10095  
10096  
10097  
10098  
10099  
100100  
100101  
100102  
100103  
100104  
100105  
100106  
100107  
100108  
100109  
100110  
100111  
100112  
100113  
100114  
100115  
100116  
100117  
100118  
100119  
100120  
100121  
100122  
100123  
100124  
100125  
100126  
100127  
100128  
100129  
100130  
100131  
100132  
100133  
100134  
100135  
100136  
100137  
100138  
100139  
100140  
100141  
100142  
100143  
100144  
100145  
100146  
100147  
100148  
100149  
100150  
100151  
100152  
100153  
100154  
100155  
100156  
100157  
100158  
100159  
100160  
100161  
100162  
100163  
100164  
100165  
100166  
100167  
100168  
100169  
100170  
100171  
100172  
100173  
100174  
100175  
100176  
100177  
100178  
100179  
100180  
100181  
100182  
100183  
100184  
100185  
100186  
100187  
100188  
100189  
100190  
100191  
100192  
100193  
100194  
100195  
100196  
100197  
100198  
100199  
100200  
100201  
100202  
100203  
100204  
100205  
100206  
100207  
100208  
100209  
100210  
100211  
100212  
100213  
100214  
100215  
100216  
100217  
100218  
100219  
100220  
100221  
100222  
100223  
100224  
100225  
100226  
100227  
100228  
100229  
100230  
100231  
100232  
100233  
100234  
100235  
100236  
100237  
100238  
100239  
100240  
100241  
100242  
100243  
100244  
100245  
100246  
100247  
100248  
100249  
100250  
100251  
100252  
100253  
100254  
100255  
100256  
100257  
100258  
100259  
100260  
100261  
100262  
100263  
100264  
100265  
100266  
100267  
100268  
100269  
100270  
100271  
100272  
100273  
100274  
100275  
100276  
100277  
100278  
100279  
100280  
100281  
100282  
100283  
100284  
100285  
100286  
100287  
100288  
100289  
100290  
100291  
100292  
100293  
100294  
100295  
100296  
100297  
100298  
100299  
100300  
100301  
100302  
100303  
100304  
100305  
100306  
100307  
100308  
100309  
100310  
100311  
100312  
100313  
100314  
100315  
100316  
100317  
100318  
100319  
100320  
100321  
100322  
100323  
100324  
100325  
100326  
100327  
100328  
100329  
100330  
100331  
100332  
100333  
100334  
100335  
100336  
100337  
100338  
100339  
100340  
100341  
100342  
100343  
100344  
100345  
100346  
100347  
100348  
100349  
100350  
100351  
100352  
100353  
100354  
100355  
100356  
100357  
100358  
100359  
100360  
100361  
100362  
100363  
100364  
100365  
100366  
100367  
100368  
100369  
100370  
100371  
100372  
100373  
100374  
100375  
100376  
100377  
100378  
100379  
100380  
100381  
100382  
100383  
100384  
100385  
100386  
100387  
100388  
100389  
100390  
100391  
100392  
100393  
100394  
100395  
100396  
100397  
100398  
100399  
100400  
100401  
100402  
100403  
100404  
100405  
100406  
100407  
100408  
100409  
100410  
100411  
100412  
100413  
100414  
100415  
100416  
100417  
100418  
100419  
100420  
100421  
100422  
100423  
100424  
100425  
100426  
100427  
100428  
100429  
100430  
100431  
100432  
100433  
100434  
100435  
100436  
100437  
100438  
100439  
100440  
100441  
100442  
100443  
100444  
100445  
100446  
100447  
100448  
100449  
100450  
100451  
100452  
100453  
100454  
100455  
100456  
100457  
100458  
100459  
100460  
100461  
100462  
100463  
100464  
100465  
100466  
100467  
100468  
100469  
100470  
100471  
100472  
100473  
100474  
100475  
100476  
100477  
100478  
100479  
100480  
100481  
100482  
100483  
100484  
100485  
100486  
100487  
100488  
100489  
100490  
100491  
100492  
100493  
100494  
100495  
100496  
100497  
100498  
100499  
100500  
100501  
100502  
100503  
100504  
100505  
100506  
100507  
100508  
100509  
100510  
100511  
100512  
100513  
100514  
100515  
100516  
100517  
100518  
100519  
100520  
100521  
100522  
100523  
100524  
100525  
100526  
100527  
100528  
100529  
100530  
100531  
100532  
100533  
100534  
100535  
100536  
100537  
100538  
100539  
100540  
100541  
100542  
100543  
100544  
100545  
100546  
100547  
100548  
100549  
100550  
100551  
100552  
100553  
100554  
100555  
100556  
100557  
100558  
100559  
100560  
100561  
100562  
100563  
100564  
100565  
100566  
100567  
100568  
100569  
100570  
100571  
100572  
100573  
100574  
100575  
100576  
100577  
100578  
100579  
100580  
100581  
100582  
100583  
100584  
100585  
100586  
100587  
100588  
100589  
100590  
100591  
100592  
100593  
100594  
100595  
100596  
100597  
100598  
100599  
100600  
100601  
100602  
100603  
100604  
100605  
100606  
100607  
100608  
100609  
100610  
100611  
100612  
100613  
100614  
100615  
100616  
100617  
100618  
100619  
100620  
100621  
100622  
100623  
100624  
100625  
100626  
100627  
100628  
100629  
100630  
100631  
100632  
100633  
100634  
100635  
100636  
100637  
100638  
100639  
100640  
100641  
100642  
100643  
100644  
100645  
100646  
100647  
100648  
100649  
100650  
100651  
100652  
100653  
100654  
100655  
100656  
100657  
100658  
100659  
100660  
100661  
100662  
100663  
100664  
100665  
100666  
100667  
100668  
100669  
100670  
100671  
100672  
100673  
100674  
100675  
100676  
100677  
100678  
100679  
100680  
100681  
100682  
100683  
100684  
100685  
100686  
100687  
100688  
100689  
100690  
100691  
100692  
100693  
100694  
100695  
100696  
100697  
100698  
100699  
100700  
100701  
100702  
100703  
100704  
100705  
100706  
100707  
100708  
100709  
100710  
100711  
100712  
100713  
100714  
100715  
100716  
100717  
100718  
100719  
100720  
100721  
100722  
100723  
100724  
100725  
100726  
100727  
100728  
100729  
100730  
100731  
100732  
100733  
100734  
100735  
100736  
100737  
100738  
100739  
100740  
100741  
100742  
100743  
100744  
100745  
100746  
100747  
100748  
100749  
100750  
100751  
100752  
100753  
100754  
100755  
100756  
100757  
100758  
100759  
100760  
100761  
100762  
100763  
100764  
100765  
100766  
100767  
100768  
100769  
100770  
100771  
100772  
100773  
100774  
100775  
100776  
100777  
100778  
100779  
100780  
100781  
100782  
100783  
100784  
100785  
100786  
100787  
100788  
100789  
100790  
100791  
100792  
100793  
100794  
100795  
100796  
100797  
100798  
100799  
100800  
100801  
100802  
100803  
100804  
100805  
100806  
100807  
100808  
100809  
100810  
100811  
100812  
100813  
100814  
100815  
100816  
100817  
100818  
100819  
100820  
100821  
100822  
100823  
100824  
100825  
100826  
100827  
100828  
100829  
100830  
100831  
100832  
100833  
100834  
100835  
100836  
100837  
100838  
100839  
100840  
100841  
100842  
100843  
100844  
100845  
100846  
100847  
100848  
100849  
100850  
100851  
100852  
100853  
100854  
100855  
100856  
100857  
100858  
100859  
100860  
100861  
100862  
100863  
100864  
100865  
100866  
100867  
100868  
100869  
100870  
100871  
100872  
100873  
100874  
100875  
100876  
100877  
100878  
100879  
100880  
100881  
100882  
100883  
100884  
100885  
100886  
100887  
100888  
100889  
100890  
100891  
100892  
100893  
100894  
100895  
100896  
100897  
100898  
100899  
100900  
100901  
100902  
100903  
100904  
100905  
100906  
100907  
100908  
100909  
100910  
100911  
100912  
100913  
100914  
100915  
100916  
100917  
100918  
100919  
100920  
100921  
100922  
100923  
100924  
100925  
100926  
100927  
100928  
100929  
100930  
100931  
100932  
100933  
100934  
100935  
100936  
100937  
100938  
100939  
100940  
100941  
100942  
100943  
100944  
100945  
100946  
100947  
100948  
100949  
100950  
100951  
100952  
100953  
100954  
100955  
100956  
100957  
100958  
100959  
100960  
100961  
100962  
100963  
100964  
100965  
100966  
100967  
100968  
100969  
100970  
100971  
100972  
100973  
100974  
100975  
100976  
100977  
100978  
100979  
100980  
100981  
100982  
100983  
100984  
100985  
100986  
100987  
100988  
100989  
100990  
100991  
100992  
100993  
100994  
100995  
100996  
100997  
100998  
100999  
1001000  
1001001  
1001002  
1001003  
1001004  
1001005  
1001006  
1001007  
1001008  
1001009  
10010010  
10010011  
10010012  
10010013  
10010014  
10010015  
10010016  
10010017  
10010018  
10010019  
100100100  
100100101  
100100102  
100100103  
100100104  
100100105  
100100106  
100100107  
100100108  
100100109  
100100110  
100100111  
100100112  
100100113  
100100114  
100100115  
100100116  
100100117  
100100118  
100100119  
100100120  
100100121  
100100122  
100100123  
100100124  
100100125  
100100126  
100100127  
100100128  
100100129  
100100130  
100100131  
100100132  
100100133  
100100134  
100100135  
100100136  
100100137  
100100138  
100100139  
100100140  
100100141  
100100142  
100100143  
100100144  
100100145  
100100146  
100100147  
100100148  
100100149  
100100150  
100100151  
100100152  
100100153  
100100154  
100100155  
100100156  
100100157  
100100158  
100100159  
100100160  
100100161  
100100162  
100100163  
100100164  
100100165  
100100166  
100100167  
100100168  
100100169  
100100170  
100100171  
100100172  
100100173  
100100174  
100100175  
100100176  
100100177  
10

648  
649

## A.3 DETAILED EXPERIMENTAL SETTINGS

650

Our training configuration and hyperparameter settings follow the default settings of the GRPO trainer under the TRL framework. For each sample, the model is prompted to generate 8 responses, each response limited to a maximum of 1024 tokens. We train on 1K samples per epoch, and to enable experiments on Qwen2.5-Math-7B, we generate one response per GPU, resulting in a total of 1K training steps. The learning rate is set to 1e-6 during training. For evaluation on the five reasoning benchmarks, all tests are conducted with a temperature setting of 0.1.

656

657

## A.4 DETAILED EXPERIMENTAL RESULTS

659

660

## A.4.1 FURTHER EXPERIMENTAL RESULTS

661

We also conduct experiments on DeepScaleR-Random-1K, with results shown in Table 3. Our method achieves performance improvements across all three different base models. Furthermore, compared to algorithm variants such as Dr.GRPO and DAPO, as well as methods that enforce reflection on the vanilla GRPO, our method demonstrates clear advantages. We also observe that the performance improvement of Qwen2.5-Math-7B is more limited compared to training on DeepScaleR-Hard-1K. We attribute this to the fact that this training data is relatively easier for the model, which further highlights the effectiveness of our method on difficult data, as such challenging data often contributes more to performance improvement.

669

670

Table 3: Pass@1 performance comparison across various mathematical evaluation benchmarks. The results below are from 1 epoch of training on **DeepScaleR-Random-1K**. The number of samples in each benchmark is indicated in parentheses. The results are evaluated under the setting of temperature = 0.1. The best results are indicated by **boldface**.

674

675

Model	Method	Avg (1560)	AIME (30)	AMC (83)	Math (500)	Min (272)	Oly (675)
Qwen2.5-Math-1.5B	Base	25.71	6.67	37.35	34.60	12.13	24.00
	SFT	30.13	10.00	30.12	47.20	14.71	24.59
	Vanilla GRPO	40.32	<b>13.33</b>	39.76	65.60	19.49	31.26
	+ Force-R	42.63	10.00	36.14	71.40	21.69	32.00
	Dr.GRPO	40.39	6.67	40.96	66.80	20.22	30.37
	DAPO	41.67	10.00	<b>44.58</b>	67.80	19.49	32.30
	EDGE-GRPO	<b>48.08</b>	13.33	<b>44.58</b>	<b>76.40</b>	<b>28.68</b>	<b>36.89</b>
Qwen2.5-Math-7B	Base	29.04	10.00	37.35	53.40	10.66	18.22
	SFT	41.99	6.67	43.37	69.00	22.06	31.41
	Vanilla GRPO	46.47	<b>23.33</b>	55.42	72.40	27.94	34.67
	+ Force-R	47.76	<b>23.33</b>	53.01	74.60	23.16	<b>38.22</b>
	Dr.GRPO	48.78	16.67	53.01	<b>75.60</b>	29.04	37.78
	DAPO	46.99	20.00	<b>59.04</b>	73.00	26.84	35.56
	EDGE-GRPO	<b>49.30</b>	16.67	50.60	<b>75.60</b>	<b>33.09</b>	37.04
Llam3.2-3B-Instruct	Base	19.81	6.67	14.46	36.20	12.13	12.00
	SFT	23.46	0.00	18.07	43.40	14.71	13.93
	Vanilla GRPO	24.49	<b>13.33</b>	18.07	45.40	15.44	13.93
	+ Force-R	23.72	10.00	16.87	44.00	12.87	14.52
	Dr.GRPO	22.89	<b>13.33</b>	<b>24.10</b>	43.20	12.50	12.30
	DAPO	21.22	6.67	9.64	40.80	13.24	12.00
	EDGE-GRPO	<b>25.90</b>	3.33	19.28	<b>47.60</b>	<b>17.28</b>	<b>15.11</b>

691

692

693

## A.4.2 PERFORMANCE COMPARISON DURING TRAINING

694

695

To visually demonstrate the superiority of our method during the training process, Figure 7 plots the performance curves of our method, the method with only forced reflection added to the vanilla GRPO (GRPO + Forced reflection), and the original GRPO method on multiple benchmark during training steps.

696

697

698

699

As can be clearly seen from the figure, our method consistently and significantly outperforms the other two baseline throughout the entire training process. In comparison, the GRPO method with only forced reflection shows some improvement over the vanilla GRPO, but the effect is limited.

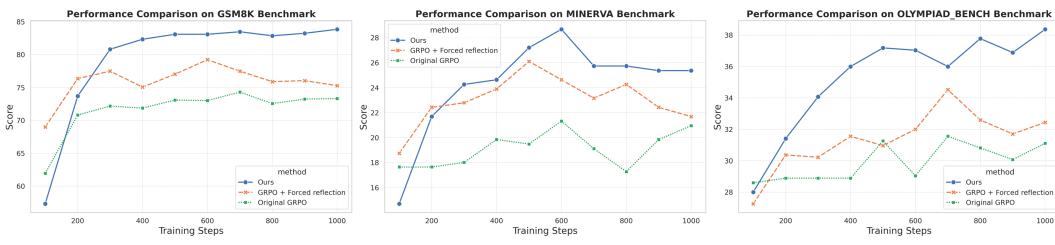


Figure 7: Performance comparison of different methods on three benchmarks during training steps. Our method consistently outperforms the vanilla GRPO and the variant with forced reflection throughout the training process.

#### A.4.3 COMPARE WITH OTHER OPEN SOURCE MODELS

To more comprehensively evaluate the effectiveness of our method, we conducted a comparison of our models against current mainstream open source models on Pass@1 performance across five mathematical reasoning benchmarks. Detailed comparison results are presented in Table 4. A core highlight is that our method achieves excellent results with extremely high data efficiency. Our models were trained using only 1K selected samples, far smaller than some other models which require tens of thousands of samples or more.

Table 4: Pass@1 performance comparison of our models against various open-source models on five mathematical reasoning benchmarks. Our models, trained on only 1K samples, demonstrate highly competitive performance. The total number of problems for each benchmark is indicated in parentheses. \* denotes data from the original paper, other results are from our own evaluation.

Model	# Train	Avg (1560)	AIME (30)	AMC (83)	Math (500)	Min (272)	Olym (675)
Qwen2.5-Math-1.5B	Base	25.71	6.67	37.35	34.60	12.13	24.00
DeepSeek-Distill-1.5B	800K	33.53	6.67	27.71	61.00	13.60	23.11
Oat-Zero-1.5B	8.5K	47.37	20.00	48.19	75.00	25.74	36.74
Qwen2.5-Math-7B	Base	29.04	10.00	37.35	53.40	10.66	18.22
DeepSeek-Distill-7B	800K	45.39	16.67	36.14	74.20	29.41	32.89
Oat-Zero-7B	8.5K	53.40	36.67	61.45	79.80	30.88	42.67
SimpleRL-Zoo-7B	8K	47.37	23.33	53.01	76.00	24.26	35.85
Eurus-7B*	48.4K	53.9	26.7	57.8	79.2	38.6	42.1
OpenReasoner-Zero-7B	5.7K	51.99	20.00	40.96	80.20	29.41	42.96
<b>EDGE-GRPO-1.5B</b>	1K	48.08	13.33	44.58	76.40	28.68	36.89
<b>EDGE-GRPO-7B</b>	1K	53.01	16.67	49.40	79.00	36.03	42.67

#### A.5 DETAILED RESULTS IN REFLECTION

This section provides a more detailed quantitative analysis of the model reflection phenomenon, which we examined from the perspectives of both spontaneous and forced reflection.

Table 5 analyzes the spontaneous reflection behavior of the models. The results show that for most models, the accuracy of responses involving spontaneous reflection is significantly lower than their overall average accuracy and the accuracy of responses without reflection. A notable exception, however, is the DeepSeek-R1-Distill series of models, which were distilled from large reasoning models. Their reflection accuracy is much higher than their average on the contrary, corroborating the point made in the main text that high-quality knowledge distillation helps improve effective reflection capabilities. Furthermore, the Qwen2.5-Math-1.5B-S1 model, trained on high-quality chain-of-thought data distilled from large reasoning models, also exhibits a reflection accuracy close to its overall accuracy, outperforming most other models.

Table 6 investigates the effect of forced reflection on the other hand. We selected samples where the models provided incorrect answers and forced them to reflect and correct their responses using four different prompts. The results show that for the vast majority of models, the improvement in accuracy from forced reflection is very limited, with correction accuracy rates generally below 10%. Even for the top-performing Deepseek-R1-Distill-Qwen-7B, the highest correction rate is only around 11%.

756 Table 5: Model performance analysis. The table compares overall accuracy, accuracy on samples  
 757 with reflection and accuracy on samples without reflection.

759 Temperature	760 Model	761 Average Acc	762 Reflection Acc	763 No-Reflection Acc
760 0.1	Qwen2.5-Math-1.5B	25.71	10.00	26.01
	DeepSeek-R1-Distill-Qwen-1.5B	33.53	42.04	14.37
	Qwen2.5-Math-1.5B-S1	27.63	27.30	27.93
	Qwen2.5-Math-1.5B-Oat-Zero	47.37	17.86	47.91
	Qwen2.5-Math-7B	29.04	13.51	29.42
	DeepSeek-R1-Distill-Qwen-7B	45.38	56.21	22.77
	Qwen2.5-Math-7B-Oat-Zero	53.40	16.00	54.01
	Qwen2.5-Math-7B-SimpleRL	47.37	25.71	47.87
	Qwen2.5-Math-7B-GPG	57.76	12.50	58.23
766 0.3	Qwen2.5-Math-1.5B	25.51	20.00	25.66
	DeepSeek-R1-Distill-Qwen-1.5B	33.27	41.47	16.44
	Qwen2.5-Math-1.5B-S1	26.86	26.73	26.97
	Qwen2.5-Math-1.5B-Oat-Zero	48.65	15.15	49.38
	Qwen2.5-Math-7B	26.28	14.81	26.48
	DeepSeek-R1-Distill-Qwen-7B	45.77	56.32	25.38
	Qwen2.5-Math-7B-Oat-Zero	52.88	16.22	53.78
	Qwen2.5-Math-7B-SimpleRL	48.14	22.86	48.72
	Qwen2.5-Math-7B-GPG	57.82	10.53	58.40
773 0.6	Qwen2.5-Math-1.5B	20.00	7.55	20.44
	DeepSeek-R1-Distill-Qwen-1.5B	34.04	41.71	22.19
	Qwen2.5-Math-1.5B-S1	20.32	24.82	16.65
	Qwen2.5-Math-1.5B-Oat-Zero	46.92	13.79	47.55
	Qwen2.5-Math-7B	20.77	9.09	21.02
	DeepSeek-R1-Distill-Qwen-7B	44.23	54.27	25.99
	Qwen2.5-Math-7B-Oat-Zero	53.08	14.29	53.97
	Qwen2.5-Math-7B-SimpleRL	45.45	17.50	46.18
	Qwen2.5-Math-7B-GPG	54.23	8.33	54.95
779 1	Qwen2.5-Math-1.5B	11.35	8.57	11.48
	DeepSeek-R1-Distill-Qwen-1.5B	27.44	40.06	16.65
	Qwen2.5-Math-1.5B-S1	20.83	21.95	19.93
	Qwen2.5-Math-1.5B-Oat-Zero	45.51	4.17	46.16
	Qwen2.5-Math-7B	16.92	5.80	17.44
	DeepSeek-R1-Distill-Qwen-7B	41.41	51.71	30.87
	Qwen2.5-Math-7B-Oat-Zero	52.50	21.62	53.25
	Qwen2.5-Math-7B-SimpleRL	45.38	20.93	46.08
	Qwen2.5-Math-7B-GPG	49.81	6.06	50.75

786  
 787 Collectively, this data indicates that a significant bottleneck persists in the self-correction capabilities  
 788 of existing models through reflection, whether spontaneous or forced. This supports the necessity of  
 789 our proposed Guided Error Correction (GEC) method.

## 792 A.6 DETAILED RESULTS IN POLICY ENTROPY

793 This section quantitatively analyzes the relationship between model confidence and answer correct-  
 794 ness using the Relative Confidence Metric (RCM). The RCM is calculated as:

$$796 \quad 797 \quad 798 \quad RCM = \frac{\text{Entropy}_{\text{Correct}} - \text{Entropy}_{\text{Incorrect}}}{\text{Average Entropy}}$$

800 As shown in Table 7, a negative RCM value indicates that the model’s correct responses have a lower  
 801 average entropy than its incorrect ones, meaning the model is more confident in its correct answers  
 802 than incorrect ones.

803 The results show that models can better calibrate their confidence on the whole, expressing higher  
 804 confidence in correct answers than in incorrect ones. However, the RCM is an aggregate metric that  
 805 reflects a macroscopic trend. The model still exhibits a significant number of high-confidence incor-  
 806 rect responses at the individual sample level. Therefore, we propose our Entropy-Driven Advantage  
 807 (EDA) to apply more fine-grained rewards and penalties at the signal level.

808 To visually substantiate our analysis of policy entropy at a fine-grained level, Figure 8 presents the  
 809 entropy distributions of correct and incorrect responses across a variety of models, including base  
 models and those enhanced through different post-training methods.

810  
 811 Table 6: Reflection accuracy under different reflection triggers. The "Incorrect" column shows the  
 812 total number of wrong answers. The subsequent columns show the reflection accuracy scores for  
 813 specific trigger words.

814 815 816 Temperature	817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 Model	840 841 842 843 844 845 Incorrect	846 847 848 849 850 851 852 Reflection accuracy (%)			
			853 854 855 Wait!	856 857 858 859 860 861 862 863 Hmm	853 854 855 Let's check it	853 854 855 Something is wrong
0.1	Qwen2.5-Math-1.5B	1159	3.624	2.675	7.161	3.365
	DeepSeek-R1-Distill-Qwen-1.5B	1037	4.638	5.700	5.507	4.251
	Qwen2.5-Math-1.5B-Oat-Zero	821	2.314	1.462	0.974	2.923
	Qwen2.5-Math-7B	1107	5.872	5.059	1.987	4.426
	DeepSeek-R1-Distill-Qwen-7B	852	8.706	11.059	8.588	8.353
	Qwen2.5-Math-7B-Oat-Zero	727	3.026	0.688	0.413	1.926
	Qwen2.5-Math-7B-SimpleRL-Zoo	821	5.366	4.146	1.098	5.122
	Qwen2.5-Math-7B-GPG	659	4.401	1.517	0.152	2.731
0.3	Qwen2.5-Math-1.5B	1162	3.184	3.356	5.594	4.389
	DeepSeek-R1-Distill-Qwen-1.5B	1041	5.967	7.507	8.277	5.101
	Qwen2.5-Math-1.5B-Oat-Zero	801	1.748	1.623	0.250	1.873
	Qwen2.5-Math-7B	1150	5.826	4.435	3.217	2.696
	DeepSeek-R1-Distill-Qwen-7B	846	10.308	9.597	8.649	9.123
	Qwen2.5-Math-7B-Oat-Zero	735	3.129	1.769	0.136	3.401
	Qwen2.5-Math-7B-SimpleRL-Zoo	809	6.057	3.585	0.865	6.057
	Qwen2.5-Math-7B-GPG	658	2.888	1.216	0.152	3.495
0.6	Qwen2.5-Math-1.5B	1248	2.648	2.809	5.056	3.852
	DeepSeek-R1-Distill-Qwen-1.5B	1029	7.101	7.879	6.323	6.323
	Qwen2.5-Math-1.5B-Oat-Zero	828	2.053	1.208	0.725	1.932
	Qwen2.5-Math-7B	1236	4.288	3.722	1.861	5.502
	DeepSeek-R1-Distill-Qwen-7B	870	11.406	12.097	11.290	10.253
	Qwen2.5-Math-7B-Oat-Zero	732	2.869	1.093	0.137	3.005
	Qwen2.5-Math-7B-SimpleRL-Zoo	851	4.935	4.113	3.055	4.465
	Qwen2.5-Math-7B-GPG	714	3.922	1.821	0.280	2.801
1	Qwen2.5-Math-1.5B	1383	1.952	1.735	2.531	1.591
	DeepSeek-R1-Distill-Qwen-1.5B	1132	9.637	10.610	9.637	8.753
	Qwen2.5-Math-1.5B-Oat-Zero	850	1.765	1.882	0.471	1.882
	Qwen2.5-Math-7B	1296	1.931	2.008	1.236	1.776
	DeepSeek-R1-Distill-Qwen-7B	914	9.430	9.649	9.320	9.539
	Qwen2.5-Math-7B-Oat-Zero	741	2.699	1.350	0.405	2.699
	Qwen2.5-Math-7B-SimpleRL-Zoo	852	3.169	2.582	0.235	2.347
	Qwen2.5-Math-7B-GPG	783	1.788	1.660	0.255	1.788

840  
 841  
 842 Each plot within the figure displays two overlapping density distributions: one for correct responses  
 843 and another for incorrect responses. These graphs visually confirm that most models exhibit mis-  
 844 judgments in their response confidence. A consistent pattern is the substantial overlap between the  
 845 entropy distributions of correct and incorrect answers.

846  
 847 Specifically, these visualizations reveal two key phenomena: Firstly, a considerable portion of in-  
 848 correct responses possesses low entropy, indicating that the models are often highly confident in  
 849

850  
 851 Table 7: Relative Confidence Metric (RCM) across different models and temperature settings. A  
 852 negative value indicates that the model exhibits lower entropy (i.e., higher confidence) in its correct  
 853 responses compared to its incorrect responses.

854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 879 Model	899 899 899 Temperature			
	899 899 899 0.1	899 899 899 0.3	899 899 899 0.6	899 899 899 1.0
Qwen2.5-Math-1.5B	0.0909	0.0278	0.0349	-0.7773
DeepSeek-R1-Distill-Qwen-1.5B	-0.1667	-0.2596	-0.2521	-0.3476
Qwen2.5-Math-1.5B-Oat-Zero	-0.5714	-0.5000	-0.5116	-0.6049
Qwen2.5-Math-7B	0.0909	-0.0294	-0.2171	-0.7175
DeepSeek-R1-Distill-Qwen-7B	-0.2222	-0.2195	-0.2406	-0.3808
Qwen2.5-Math-7B-Oat-Zero	-0.6667	-0.6250	-0.5625	-0.5556
Qwen2.5-Math-7B-SimpleRL	-0.2727	-0.2813	-0.3553	-0.5179
Qwen2.5-Math-7B-GPG	-0.6667	-0.5000	-0.7059	-0.7778

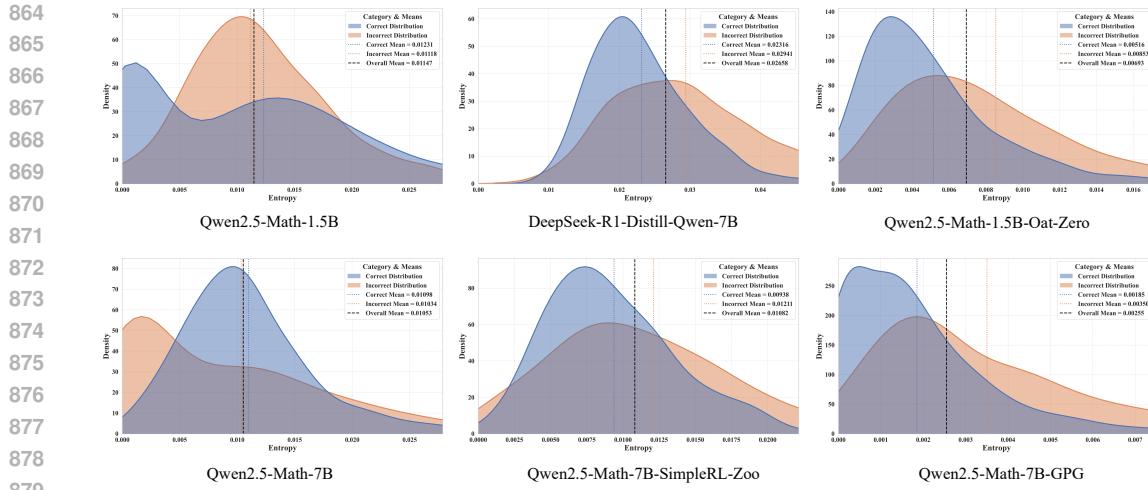


Figure 8: The entropy distribution of correct and incorrect responses within different models. The results are evaluated under the setting of temperature = 0.1.

their erroneous answers. Secondly, many correct responses exhibit high entropy, suggesting a lack of confidence even when the model produces the right answer.

This evidence highlights that relying on aggregate metrics like average entropy is insufficient, as it masks these critical sample-level discrepancies. The observed miscalibration of confidence at this granular level strongly motivates our proposed Entropy-Driven Advantage (EDA) mechanism, which is designed to apply more precise rewards and penalties to address these confidence misjudgments directly.

#### A.7 USE OF LARGE LANGUAGE MODELS

Some portions of the text were polished with the assistance of Large Language Models (LLMs). All content remains the responsibility of the authors.

#### A.8 ETHICS STATEMENT

This study does not involve human subjects, sensitive data, or potentially harmful applications. All mathematical reasoning datasets used (e.g., DeepScaleR, MATH, AIME) are publicly available and contain no personal privacy information or copyright-protected content. The authors have read and pledged to uphold the ICLR Code of Ethics. The research encountered no conflicts of interest, discriminatory bias, or legal-compliance risks. Every experiment followed academic integrity principles, and all reported results are truthful and unaltered.

#### A.9 REPRODUCIBILITY STATEMENT

Appendix A.2 presents the complete algorithmic pseudocode, Appendix A.3 lists all hyperparameters and training schedules, and Section 5 describes the data filtering pipeline in detail. Following the prescriptions in these sections enables full replication of the reported results.