Medium-GRPO: Online Difficulty-Adaptive Calibration for Enhancing LLM Reasoning

Yihang Feng[1], Fei Su[1], Zhe Cui[1,2,*]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, Beijing, China

***Abstract*—Group Relative Policy Optimization (GRPO) is widely adopted in the post-training of Large Language Models (LLMs) to enhance mathematical reasoning performance. However, its effectiveness depends critically on the variance of rewards within a group of sampled responses. When the difficulty of the dataset is misaligned with the model's capability, the model often yields groups of responses that are uniformly correct or incorrect, leading to vanishing gradient signals that hinder optimization. Addressing the limitations of existing augmentation methods, we propose Medium-GRPO, an online difficulty-adaptive calibration framework. Leveraging the model's intrinsic capabilities, Medium-GRPO utilizes an expression-based strategy to shift problems that are overly hard or easy for the model into the medium-difficulty range. We further introduce a fine-grained difficulty metric based on response length statistics to improve modification success rates, and integrate the modification process into the training objective for joint optimization. Experiments on five mathematical benchmarks show that Medium-GRPO consistently outperforms GRPO. Notably, on the challenging AIME-24 and AIME-25 benchmarks, our method achieves relative performance gains of 16.4% and 11.2% with Qwen3-4B.**

***Index Terms*—Reinforcement Learning, Large Language Models, data augmentation, mathematical reasoning**

I. Introduction

Reinforcement Learning (RL) has emerged as a paramount paradigm in the post-training phase of Large Language Models (LLMs) [1]–[3], significantly enhancing reasoning capabilities, particularly in complex tasks such as mathematical reasoning and code generation. Among these approaches, Group Relative Policy Optimization (GRPO) [4] is widely favored for its efficiency in optimizing models by integrating verifiable rewards [2], [5]. However, the core mechanism of GRPO involves sampling a group of responses for each query and estimating advantages based on the relative differences in rewards within the group. While this mechanism eliminates the need for an additional critic model, it imposes stringent requirements on the training dataset: when the data difficulty is 'too easy' or 'too difficult' relative to the model's current capability, a substantial proportion of the dataset tends to yield homogenized generations, where the responses within a group are either entirely correct or entirely incorrect, referred to as zero-variance prompts [6]. For these instances, the relative advantage becomes zero, resulting in vanishing gradient signals that make no substantive contribution to the optimization of the policy model. Consequently, achieving dynamic alignment between training data difficulty and model capability to efficiently leverage existing datasets is crucial for advancing GRPO performance.

Existing data augmentation methods exhibit significant limitations in addressing this challenge. Mainstream approaches [7]–[9] primarily target the Supervised Fine-Tuning (SFT) phase, focusing on statically expanding data scale and diversity prior to training without accounting for the model's real-time performance, rendering them ill-suited for the GRPO framework. While recent methods such as Guide-GRPO [10] and QUESTA [11] attempt to introduce prompts during the RL process to assist with hard problems, they typically rely on stronger teacher models or human annotation. This not only incurs additional data costs but also neglects the handling of overly easy problems. In contrast, works like SPIN [12] and SvS [13] introduce the concept of Self-Play into LLM post-training, utilizing the model itself to generate data while simultaneously training both solving and synthesis capabilities. This offers pivotal inspiration for resolving the data alignment problem via the model's endogenous capabilities.

To this end, we propose Medium-GRPO, a difficulty-adaptive calibration framework for online data augmentation. Specifically, we design an expression-based difficulty modification strategy in which the policy model modifies problem difficulty while maintaining answer invariance. This approach not only improves the modification success rate but also preserves the distributional and stylistic integrity of the original dataset without external annotation. To efficiently target candidates suitable for modification, we introduce a fine-grained difficulty metric based on response length statistics. Moreover, the difficulty adjustment process itself is integrated into the training objective; the average accuracy of the group responses to the modified problems serves as a metric for modification success, thereby achieving the co-evolution of the model's problem-solving and adaptive modification capabilities.

We evaluate our method across five widely used mathematical benchmarks with diverse difficulty distributions. Experimental results demonstrate that our method consistently outperforms the conventional GRPO baseline on the Mean@32 metric. The efficacy of our approach is particularly pronounced on complex tasks, where our method achieves relative improvements of 16.4% on AIME-24 and 11.2% on AIME-25 with Qwen3-4B. These results validate that utilizing the model's intrinsic capacity for difficulty calibration is a scalable and effective path to unlock the full potential of RL post-training. Code is available at github.com/flyhigh77/MGRPO.

*Corresponding author: cuizhe@bupt.edu.cn

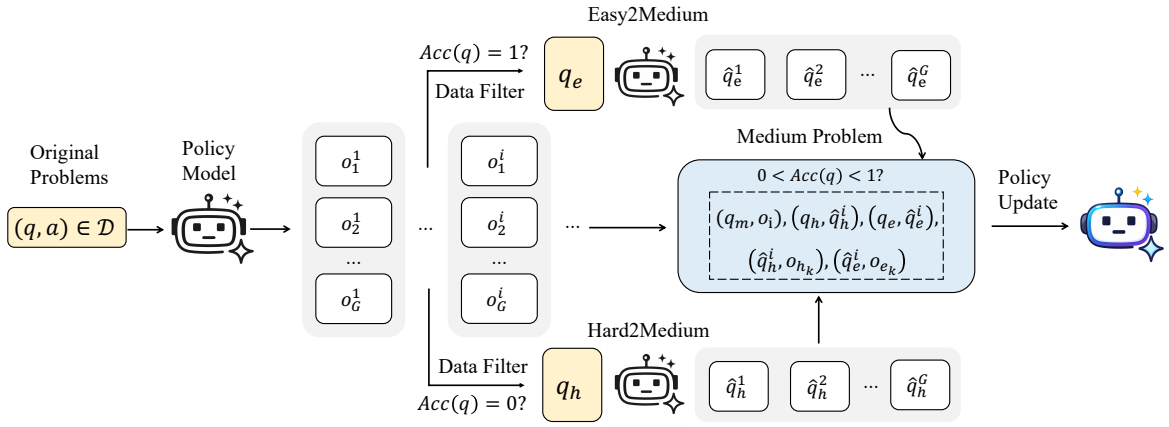

Fig. 1: Overview of the Medium-GRPO framework. Easy and hard problems undergo data filtering and are subsequently transformed into medium instances via the *Easy-to-Medium* and *Hard-to-Medium* branches. The original medium-difficulty samples, modification tasks, and modified problems are jointly utilized to update the policy model.

II. Method

A. Overview

GRPO estimates the advantage in a group-relative manner. For each query q, it samples a group of G responses $\{o_i\}_{i=1}^G$ and computes the advantage of the i-th response by normalizing the corresponding group-level rewards $\{R_i\}_{i=1}^G$:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^G)}{\text{std}(\{R_j\}_{j=1}^G)} \tag{1}$$

When the model encounters problems that are either overly difficult or overly easy, the responses within a group tend to yield uniform outcomes (i.e., all correct or all incorrect), causing the advantage signal to degrade to zero. As a result, such samples fail to contribute effective gradient information for policy updates. These samples often constitute a significant proportion of the data when there is a misalignment between dataset difficulty and model capability.

To fully exploit the potential value of these samples, we propose Medium-GRPO, a data augmentation method based on difficulty adaptive calibration. As illustrated in Fig. 1, during training, given a problem–answer pair (q, a) from the dataset $\mathcal{D}$, the policy model generates G responses $\{o_i\}_{i=1}^G$. We identify samples whose group accuracy is exactly 0 (hard problems, denoted by q_h) or 1 (easy problems, denoted by q_e), and select a subset of them that exhibit potential for modification.

Through two distinct branches—*Easy-to-Medium* and *Hard-to-Medium*—we then adjust the difficulty of q_e and q_h, respectively, shifting them into a moderate difficulty regime that provides more informative advantage signals. Inspired by the self-play paradigm introduced in prior work, the difficulty modification task is autonomously performed by the current policy model, and can be written as:

$$\{\hat{q}_h^i\}_{i=1}^G \sim \pi_\theta(\cdot \mid \text{h2m}(q_h)), \{\hat{q}_e^i\}_{i=1}^G \sim \pi_\theta(\cdot \mid \text{e2m}(q_e)). \tag{2}$$

The final training batch thus consists of three components: (1) Solutions to the original medium-difficulty problems:(q_m, o_i); (2) The difficulty modification tasks: $(q_h, \hat{q}_h^i)$, $(q_e, \hat{q}_e^i)$; (3) Solutions to the successfully calibrated problems:$(\hat{q}_h^i, o_{h_k})$, $(\hat{q}_e^i, o_{e_k})$. These data jointly contribute to policy updates, enhancing the density of advantage signals and the stability of gradients throughout training.

B. Expression-based Difficulty Modification

For overly difficult problems, a straightforward approach to reducing difficulty is to introduce additional information to assist the model in solving the problem [14]. However, this approach faces significant challenges. Merely providing brief natural-language hints typically yields negligible improvements. Conversely, incorporating verbose descriptions tends to disrupt the distributional homogeneity of the original dataset and bias the model's reasoning patterns [15]. Furthermore, directly rewriting the problem statement is equally problematic, as it is challenging to guarantee the invariance of the ground-truth answers.

Motivated by these considerations, we propose an expression-based difficulty modification strategy. Specifically, we provide the policy model with both the hard problem and its ground-truth answer, prompting it to perform reverse reasoning starting from the answer. The goal is to extract the single most pivotal mathematical expression—whether an equation, inequality, or function—required for the solution. Subsequently, this derived expression is incorporated into the problem statements as a hint. This approach confines the auxiliary information to a valid scope, thereby enhancing its effectiveness. Furthermore, by incorporating strictly mathematical symbols without extraneous natural-language descriptions, our method ensures that the new problems maintain both distributional homogeneity and stylistic consistency with the original dataset.

For overly simple problems, we adopt an analogous methodology. Given the original problem and its answer, the policy

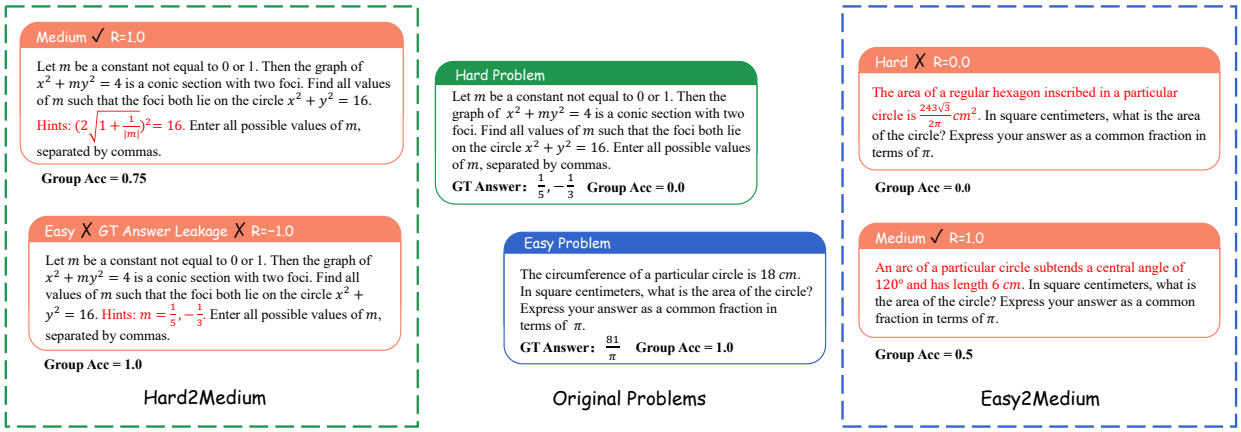

Fig. 2: Illustrative examples of the proposed expression-based difficulty modification strategy and the reward assignment.

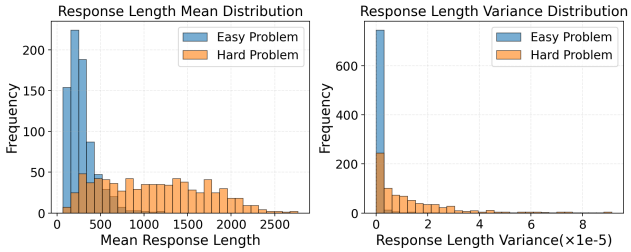

Fig. 3: Distribution analysis of response length statistics. Hard problems exhibit significantly larger means and variances compared to easy problems.

model is prompted to identify the most obvious mathematical condition in the problem statement. While keeping the ground-truth answer unchanged, the model is instructed to employ reverse reasoning to substitute this condition with a more implicit one, which necessitates a more complex reasoning path and the application of additional mathematical knowledge to resolve the problem, thereby elevating the problem difficulty with minimal modifications to the problem statements.

C. Difficulty-Aware Reward Formulation

The difficulty modification task is integrated into the reinforcement learning framework. For each hard problem q_h and the easy problem q_e, we generate G modified problems and evaluate the policy model's performance on these variations. A positive reward ($r = 1$) is assigned if the group accuracy of the modified problem falls strictly within the interval $(0, 1)$, indicating successful difficulty modulation. Conversely, an accuracy of exactly 0 or 1 indicates an ineffective modification, as the modified problem remains overly difficult or overly easy, and results in a zero reward($r = 0$). Furthermore, to prevent degenerate strategies such as directly leaking the ground-truth answer in the problem statement or failing to perform any substantive modification, we also introduce an explicit penalty mechanism with $r = -1$.

An example of difficulty modification and reward assignment is shown in Fig. 2. In the Hard-to-Medium case, the original conic section problem is initially unsolvable with a group accuracy of 0. Our method extracts a pivotal intermediate equation, $(2\sqrt{1+\frac{1}{|m|}})^2 = 16$, as a hint. This reasonable adjustment calibrates the group accuracy to 0.75, thereby earning a positive reward ($r = 1.0$). Conversely, a degenerate modification that directly leaks the ground-truth answer triggers our penalty mechanism ($r = -1.0$). In the Easy-to-Medium case, the original problem has a group accuracy of 1.0. By substituting the explicit circumference condition with an implicit description of an arc, the accuracy is successfully adjusted to the target level of 0.5, transforming the sample into a valid training instance ($r = 1.0$). However, an overly aggressive rewrite introducing the concept of "inscribed regular hexagon" renders the modified problem excessively difficult, dropping the accuracy to 0.0 and resulting in a failed difficulty modification ($r = 0.0$).

D. Data Filtering Strategy

The methods described above are still inherently constrained by the capability of the policy model. Small language models may lack the capacity to effectively calibrate the majority of the dataset. Consequently, indiscriminately attempting to modify all hard and easy problems would not only substantially increase training time, but also introduce excessive noise due to a large number of low-quality modification attempts, thereby interfering with the model's primary objective of mathematical problem solving.

Therefore, we selectively apply difficulty modification only to the easiest subset among hard problems and the hardest subset among easy problems. When problems share identical accuracy outcomes, we further measure difficulty using the mean and variance of the lengths of the group responses $\{o_i\}_{i=1}^{G}$.As illustrated in Fig. 3, we analyze the distributions of these statistics for hard and easy problems, and observe that hard problems exhibit significantly larger mean response lengths and variances than easy ones.

Based on this observation, we design a data filtering strategy as follows. Within each training batch, we first identify the sets of hard and easy problems. For each problem in either set,

Model	Method	AIME-24	AIME-25	MATH-500	AMC23	BeyondAIME
Qwen3-4B	Init Model	22.60	18.44	79.62	66.17	9.88
	GRPO	32.40	25.00	85.04	75.31	15.59
	Ours	37.71	27.80	85.49	76.33	17.28
	Δ	+16.4%	+11.2%	+0.5%	+1.4%	+10.8%
Qwen3-1.7B	Init Model	13.44	10.52	69.38	44.53	4.28
	GRPO	18.85	18.75	77.11	55.08	6.66
	Ours	19.79	19.48	76.96	56.09	6.88
	Δ	+5.0%	+3.9%	-0.2%	+1.8%	+3.3%

TABLE I: Performance comparison across five mathematical reasoning benchmarks. Δ denotes the relative improvement of ours method over GRPO.

we compute the problem-level mean and variance of response lengths, denoted as $\mu(o)$ and $\nu(o)$, and apply normalization within each set:

$$z_\mu(o) = \frac{\mu(o) - \bar{\mu}}{\sigma_\mu + \epsilon}, \quad z_\nu(o) = \frac{\nu(o) - \bar{\nu}}{\sigma_\nu + \epsilon}, \tag{3}$$

where $\bar{\mu}$ and σ_μ denote the set-level mean and standard deviation of $\mu(o)$, and $\bar{\nu}$ and σ_ν denote those of $\nu(o)$. The final composite difficulty metric is defined as:

$$\text{score}(x) = z_\mu(x) + \alpha \cdot z_\nu(x), \tag{4}$$

where α is a hyperparameter governing the trade-off between the mean and variance. A higher score indicates a higher difficulty level. Based on this metric, we select only the proportion θ of the sets for modification.

Moreover, given that we generate G candidate variations for each problem, incorporating all variants into the training batch would introduce unnecessary redundancy. We consider the expectation of the absolute advantage, let p denote the accuracy of a problem, and let $\hat{A}_{\text{pos}}$ and $\hat{A}_{\text{neg}}$ represent the advantages of correct and incorrect responses, respectively:

$$\begin{aligned} \mathbb{E}\big[\,|\hat{A}|\,\big] &= p \cdot |\hat{A}_{\text{pos}}| + (1-p) \cdot |\hat{A}_{\text{neg}}| \\ &= p \cdot \frac{1-p}{\sqrt{p(1-p)}} + (1-p) \cdot \frac{p}{\sqrt{p(1-p)}} \\ &= 2\sqrt{p(1-p)} \end{aligned} \tag{5}$$

We observe that the expected absolute advantage is maximized when the accuracy $p = 0.5$. Consequently, from the pool of all modified versions for a given problem, we select only the single candidate with an accuracy closest to 0.5 for inclusion in the training set.

III. Experiments

A. Settings

To validate our approach, we adopt Qwen3-1.7B [16] and Qwen3-4B [16] as our base models. MATH-12K [17] is used as the training dataset.For evaluation, we selected five benchmarks exhibiting diverse difficulty distributions: AIME-24 [18], AIME-25 [18], MATH-500 [19], AMC23 [20], and BeyondAIME [21].

The model operates in non-thinking mode. During training, we sample 8 responses per problem with Top-p = 1.0. The models are trained for three epochs with a global batch size of 128. For evaluation, we perform 32 independent runs for each problem to reduce variance and improve result stability. The final performance is reported as the average Pass@1 score, denoted as Mean@32. The temperature is fixed to 1.0. To ensure a fair comparison, we restrict modifications to only $\theta_h = 10\%$ of hard problems and $\theta_e = 5\%$ of easy problems, ensuring our training duration does not significantly exceed that of the GRPO baseline. The hyperparameter α in Equation (4) is set to 0.3. All experiments are implemented using the VeRL [22] framework, utilizing vLLM [23] for inference acceleration.

B. Main Results

We present the main experimental results in TABLE I. Overall, our proposed method consistently outperforms the GRPO baseline across the majority of benchmarks and model scales. The most significant improvements are observed on challenging competition-level benchmarks, such as AIME-24, and AIME-25. For the Qwen3-4B model, our method achieves a remarkable boost, surpassing GRPO by 16.4% on AIME-24 and 11.2% on AIME-25. This suggests that the problem difficulty modification task, which necessitates reverse reasoning conditioned on the ground truth, effectively bolsters the model's reasoning capabilities for complex problems. In contrast, the conventional GRPO baseline exhibits limited improvement on challenging benchmarks, likely because the MATH-12K training set is predominantly composed of easier problems. Notably, on benchmarks where the difficulty

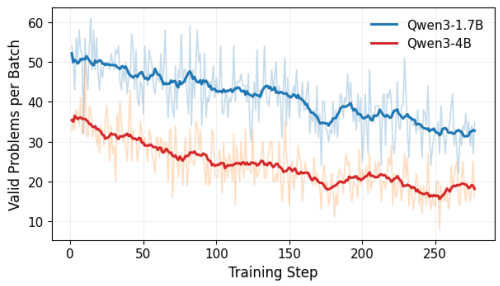

Fig. 4: The number of valid samples from the original dataset within a single training batch.

TABLE II: Ablation study on the contribution of difficulty modification branches. Both the Easy-to-Medium branch and the subsequent integration of the Hard-to-Medium branch contribute to the performance gains over the GRPO baseline.

Model	Method	AIME-24	AIME-25	MATH-500	AMC23	BeyondAIME
Qwen3-4B	GRPO	32.40	25.00	85.04	75.31	15.59
	+E2M	36.67	**28.33**	84.89	**77.34**	16.16
	+H2M	**37.71**	27.80	**85.49**	76.33	**17.28**

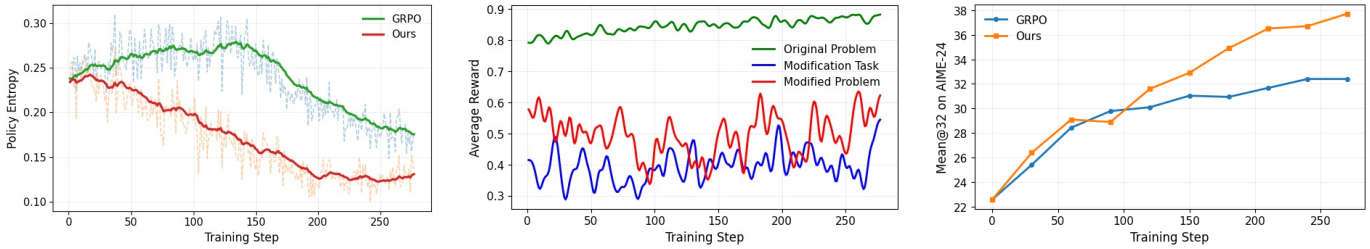

Fig. 5: Analysis of learning dynamics. (Left) Comparison of policy entropy curves between GRPO and our method. (Center) Average reward trajectories for distinct data components. (Right) Performance evaluation on AIME-24.

TABLE III: Ablation study on the expression-based difficulty modification strategy.

(a) Easy-to-Medium($N = 2000$)

Difficulty	Original	Free-Form	Expression
Hard	0	523	619
Medium	0	404	**455**
Easy	2000	1073	926

(b) Hard-to-Medium($N = 767$)

Difficulty	Original	Free-Form	Expression
Hard	767	697	453
Medium	0	62	**194**
Easy	0	8	120

distribution aligns more closely with the training data—such as AMC23 and MATH-500—our method yields performance gains comparable to those of GRPO.

Validating the effectiveness of our approach across different model scales, we observe a positive correlation between model capacity and performance gains. The Qwen3-4B model benefits more substantially from our training strategy than the Qwen3-1.7B model. For example, the relative gain on AIME-24 increases from 5.0% (1.7B) to 16.4% (4B). This phenomenon is primarily attributed to the fact that the efficacy of difficulty adjustment is intrinsically constrained by the base model's capabilities. Compared to Qwen3-1.7B, the stronger Qwen3-4B is capable of generating higher-quality and more modifications across a broader range of hard and easy instances, thereby driving more significant performance gains. Conversely, the intrinsic difficulty distribution of MATH12K is more congruent with the capacity of Qwen3-1.7B. As illustrated in Fig. 4, the original training data already provides substantial effective gradients for the 1.7B model, which consequently diminishes the relative impact of the modified problems.

C. Learning Dynamics

We tracked the average performance metrics on AIME-24 and AIME-25 across varying steps throughout the training process, as illustrated in Fig. 5(right). The results demonstrate that, overall, our method consistently outperforms the conventional GRPO baseline.

To understand the underlying mechanism, we further analyze the evolution of the model's policy entropy in Fig. 5(left). By modifying both hard and easy instances, Medium-GRPO provides richer and more effective gradient information at each training step. Consequently, as the curves indicate, the policy entropy of Medium-GRPO exhibits a significantly more stable decline during the initial phase compared to that of GRPO, thereby leading to superior performance gains. A potential concern is that the eventual collapse of policy entropy could detrimentally impact model performance as training progresses [24], [25]. However, the trajectory reveals that while the entropy stabilizes during the final 100 steps, the performance continues to improve. This resilience is attributed to the difficulty modification tasks, which play a crucial role in sustaining policy entropy. As shown in Fig. 5 (Center), the average reward statistics indicate that both the difficulty modification tasks and the modified problems pose significantly greater challenges to the model compared to the original dataset. The model's performance on these tasks remains far from saturation, thereby effectively preventing the indefinite collapse of policy entropy.

D. Ablation Studies

We first analyze how different difficulty modification sources contribute to the downstream reasoning performance of the Qwen3-4B model. We trained Qwen3-4B using solely the Easy-to-Medium branch under identical settings. As shown in TABLE II, the results indicate that both the Easy-to-Medium

and Hard-to-Medium branches contribute effectively to the model's performance improvement.

To validate the effectiveness of expression-based difficulty modification strategy, we selected all hard problems and a subset of easy problems identified for Qwen3-4B. We instructed Qwen3-4B to modify these instances and subsequently evaluated its performance on the modified problems to monitor shifts in the difficulty distribution. As a control baseline, we maintained all other requirements invariant but removed the specific constraints regarding mathematical expressions from the original prompt. Specifically, for hard problems, the model was permitted to append arbitrary additional information; for easy problems, it was granted unrestricted freedom to rewrite the problem statements. As shown in TABLE III, compared to the unconstrained baseline, our approach successfully converts substantially more instances into the target 'Medium' category. This demonstrates that the expression-based difficulty modification strategy effectively constrains the model's modification space, thereby significantly improving the success rate of the task.

IV. Conclusion

In this work, we addressed the vanishing advantage problem in GRPO training caused by the misalignment between dataset difficulty and model capability. We proposed Medium-GRPO, an online data augmentation framework that performs adaptive difficulty calibration. By introducing an expression-based modification strategy coupled with a fine-grained filtering mechanism, we successfully guided the policy model to autonomously convert "zero-variance" problems into informative instances rich in gradient signals. Empirical results confirm that our method significantly enhances performance on complex mathematical reasoning tasks without requiring external annotation, exhibiting particularly superior robustness on high-difficulty benchmarks. Our work not only alleviates the stringent data quality requirements of GRPO but also offers a promising avenue for leveraging the endogenous capabilities of LLMs to achieve data-efficient self-evolution.

Acknowledgment

This work is supported by Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

References

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[2] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[3] K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao *et al.*, "Kimi k1. 5: Scaling reinforcement learning with llms," *arXiv preprint arXiv:2501.12599*, 2025.

[4] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[5] N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu *et al.*, "Tulu 3: Pushing frontiers in open language model post-training," *arXiv preprint arXiv:2411.15124*, 2024.

[6] H. Zheng, Y. Zhou, B. R. Bartoldson, B. Kailkhura, F. Lai, J. Zhao, and B. Chen, "Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts," *arXiv preprint arXiv:2506.02177*, 2025.

[7] L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu, "Metamath: Bootstrap your own mathematical questions for large language models," *arXiv preprint arXiv:2309.12284*, 2023.

[8] Z. Tang, X. Zhang, B. Wang, and F. Wei, "Mathscale: Scaling instruction tuning for mathematical reasoning," *arXiv preprint arXiv:2403.02884*, 2024.

[9] Q. Pei, L. Wu, Z. Pan, Y. Li, H. Lin, C. Ming, X. Gao, C. He, and R. Yan, "Mathfusion: Enhancing mathematical problem-solving of llm through instruction fusion," *arXiv preprint arXiv:2503.16212*, 2025.

[10] V. Nath, E. Lau, A. Gunjal, M. Sharma, N. Baharte, and S. Hendryx, "Adaptive guidance accelerates reinforcement learning of reasoning models," *arXiv preprint arXiv:2506.13923*, 2025.

[11] J. Li, H. Lin, H. Lu, K. Wen, Z. Yang, J. Gao, Y. Wu, and J. Zhang, "Questa: Expanding reasoning capacity in llms via question augmentation," *arXiv preprint arXiv:2507.13266*, 2025.

[12] Z. Chen, Y. Deng, H. Yuan, K. Ji, and Q. Gu, "Self-play fine-tuning converts weak language models to strong language models," *arXiv preprint arXiv:2401.01335*, 2024.

[13] X. Liang, Z. Li, Y. Gong, Y. Shen, Y. N. Wu, Z. Guo, and W. Chen, "Beyond pass@ 1: Self-play with variational problem synthesis sustains rlvr," *arXiv preprint arXiv:2508.14029*, 2025.

[14] V. Agrawal, P. Singla, A. S. Miglani, S. Garg, and A. Mangal, "Give me a hint: Can llms take a hint to solve math problems?" *arXiv preprint arXiv:2410.05915*, 2024.

[15] W. Hua, K. Zhu, L. Li, L. Fan, M. Jin, S. Lin, H. Xue, Z. Li, J. Wang, and Y. Zhang, "Disentangling logic: The role of context in large language model reasoning capabilities," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 19 219–19 242.

[16] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[17] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," *arXiv preprint arXiv:2103.03874*, 2021.

[18] Mathematical Association of America, "American Invitational Mathematics Examination (AIME)," https://maa.org/math-competitions/aime, 2025, [Accessed: Oct. 2025].

[19] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," in *The Twelfth International Conference on Learning Representations*, 2023.

[20] Mathematical Association of America, "American Mathematics Competitions (AMC 10/12)," https://maa.org/math-competitions/amc, 2023, [Accessed: Oct. 2025].

[21] ByteDance-Seed, "BeyondAIME: Advancing Math Reasoning Evaluation Beyond High School Olympiads," https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME, 2025, [Accessed: Oct. 2025].

[22] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu, "Hybridflow: A flexible and efficient rlhf framework," in *Proceedings of the Twentieth European Conference on Computer Systems*, 2025, pp. 1279–1297.

[23] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the 29th symposium on operating systems principles*, 2023, pp. 611–626.

[24] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu *et al.*, "Dapo: An open-source llm reinforcement learning system at scale," *arXiv preprint arXiv:2503.14476*, 2025.

[25] G. Cui, Y. Zhang, J. Chen, L. Yuan, Z. Wang, Y. Zuo, H. Li, Y. Fan, H. Chen, W. Chen *et al.*, "The entropy mechanism of reinforcement learning for reasoning language models," *arXiv preprint arXiv:2505.22617*, 2025.