Harder Is Better: Boosting Mathematical Reasoning via Difficulty-Aware GRPO and Multi-Aspect Question Reformulation

Yanqi Dai (代彦琪)[1,2], Yuxiang Ji[2,3], Xiao Zhang[2,4], Yong Wang[2]*, Xiangxiang Chu[2], Zhiwu Lu[1]*

[1]Gaoling School of Artificial Intelligence, Renmin University of China

[2]AMAP, Alibaba Group [3]Xiamen University [4]Dalian University of Technology

Motivation & Preliminaries

Motivation: Reinforcement Learning with Verifiable Rewards (RLVR)

- **[Algorithm] GRPO:** implicitly weaken the update magnitude of harder questions.
- **[Data] WizardMath, MetaMath:** need to generate both questions and answers from scratch, or just rephrase questions to enhance diversity rather than difficulty.

Preliminaries: Optimization Objective of Group Relative Policy Optimization (GRPO)

Specifically, GRPO optimizes the policy model π_θ by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[q \sim \mathcal{D}, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)\right]$$

$$\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\left\{\min\left[I_{it}(\theta)\hat{A}_{\text{GR},i}, \text{clip}\left(I_{it}(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{\text{GR},i}\right]\right\}, \quad (1)$$

$$\text{where } I_{it}(\theta) = \frac{\pi_\theta\left(o_{i,t} \mid q, o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t} \mid q, o_{i,<t}\right)}, \ \hat{A}_{\text{GR},i} = \frac{r_i - \text{mean}\left(\{r_i\}_{i=1}^{G}\right)}{\text{std}\left(\{r_i\}_{i=1}^{G}\right)}. \quad (2)$$

Optimization Objective of Difficulty-Aware Group Policy Optimization (DGPO)

Specifically, the optimization objective of DGPO is defined as follows:

$$\mathcal{J}_{\text{DGPO}}(\theta) = \mathbb{E}\left[\{q_s\}_{s=1}^{B} \sim \mathcal{D}, \{o_{si}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q_s)\right]$$

$$\frac{1}{\sum_{s=1}^{B_{\text{v}}} \sum_{i=1}^{G} |o_{si}|} \sum_{s=1}^{B_{\text{v}}} \lambda_s \sum_{i=1}^{G} \sum_{t=1}^{|o_{si}|} \left\{ \min\left[I_{sit}(\theta) \hat{A}_{\text{DG},si}, \text{clip}\left(I_{sit}(\theta), 1-\varepsilon, 1+\varepsilon\right) \hat{A}_{\text{DG},si} \right] \right\}, \quad (3)$$

where $I_{sit}(\theta)$ is the importance sampling ratio of the token $o_{si,t}$, and $\hat{A}_{\text{DG},si}$ is the advantage of the response o_i obtained by DGAE, respectively given by:

$$I_{sit}(\theta) = \frac{\pi_\theta\left(o_{si,t} \mid q_s, o_{si,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{si,t} \mid q_s, o_{si,<t}\right)}, \quad \hat{A}_{\text{DG},si} = \frac{r_{si} - \text{mean}\left(\{r_{si}\}_{i=1}^{G}\right)}{\text{MAD}\left(\{r_{si}\}_{i=1}^{G}\right)}, \quad (4)$$

$$\text{where } \text{MAD}\left(\{r_{si}\}_{i=1}^{G}\right) = \frac{1}{G} \sum_{i=1}^{G} \left| r_{si} - \text{mean}\left(\{r_{si}\}_{i=1}^{G}\right) \right|. \quad (5)$$

Here, $\text{MAD}(\cdot)$ denotes the mean absolute deviation function. Furthermore, λ_s is the difficulty-aware weight for the query q_s computed by DQW as follows:

$$\lambda_s = B_{\text{v}} \cdot \frac{\exp\left(D_s/T\right)}{\sum_{s=1}^{B_{\text{v}}} \exp\left(D_s/T\right)}, \quad \text{where } D_s = -\text{mean}\left(\{r_{si}\}_{i=1}^{G}\right). \quad (6)$$

Optimization Objective of Difficulty-Aware Group Policy Optimization (DGPO)

Difficulty-balanced Group Advantage Estimation (DGAE)

Consider a single question q and its corresponding responses $\{o_i\}_{i=1}^{G}$, the unclipped policy gradient calculated in GRPO is as follows:

$$g_{\text{GRPO}} = \frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\hat{A}_{\text{GR},i}\nabla_\theta I_{it}(\theta)$$

$$= \frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\underbrace{\text{sgn}\left(\hat{A}_{\text{GR},i}\right)}_{\text{update direction}}\underbrace{\left|\hat{A}_{\text{GR},i}\right|}_{\text{update magnitude}}\underbrace{\text{detach}\left(I_{it}(\theta)\right)}_{\text{importance sampling ratio}}\underbrace{\nabla_\theta \log\left(\pi_\theta\left(o_{i,t}\mid q, o_{i,<t}\right)\right)}_{\text{likelihood gradient}}, \quad (7)$$

Theorem 1 (Update Magnitude for a Single Question using GRAE)**.** *Given a single question q and its corresponding responses $\{o_i\}_{i=1}^{G}$, each query-response pair receives a binary accuracy reward $r_i \in \{0,1\}$, and p represents the accuracy rate, i.e., the proportion for a reward of 1. Then, the total update magnitude without clipping for the single question q when using GRAE satisfies:*

$$\sum_{i=1}^{G}\left|\hat{A}_{\text{GR},i}\right| = \sum_{i=1}^{G}\left|\frac{r_i - \text{mean}\left(\{r_i\}_{i=1}^{G}\right)}{\text{std}\left(\{r_i\}_{i=1}^{G}\right)}\right| = 2G\sqrt{p(1-p)}, \text{ where } p = \frac{1}{G}\sum_{i=1}^{G} r_i. \quad (8)$$

This total update magnitude varies with respect to the accuracy rate p, reaching its maximum when $p = 0.5$ and gradually decreasing as p approaches either 0 or 1.

Optimization Objective of Difficulty-Aware Group Policy Optimization (DGPO)

Difficulty-balanced Group Advantage Estimation (DGAE)

Specifically, the advantage function of DGAE is defined as follows:

$$\hat{A}_{\mathrm{DG},i} = \frac{r_i - \mathrm{mean}\left(\{r_i\}_{i=1}^{G}\right)}{\mathrm{MAD}\left(\{r_i\}_{i=1}^{G}\right)}, \quad \text{where } \mathrm{MAD}\left(\{r_i\}_{i=1}^{G}\right) = \frac{1}{G}\sum_{i=1}^{G}\left|r_i - \mathrm{mean}\left(\{r_i\}_{i=1}^{G}\right)\right|. \quad (9)$$

Theorem 2 (Update Magnitude for a Single Question using DGAE)**.** *Given a single question q and its corresponding responses $\{o_i\}_{i=1}^{G}$, each query-response pair receives a reward r_i. Then, the total update magnitude without clipping for the single question q when using DGAE satisfies:*

$$\sum_{i=1}^{G}\left|\hat{A}_{\mathrm{DG},i}\right| = \sum_{i=1}^{G}\left|\frac{r_i - \mathrm{mean}\left(\{r_i\}_{i=1}^{G}\right)}{\frac{1}{G}\sum_{i=1}^{G}\left|r_i - \mathrm{mean}\left(\{r_i\}_{i=1}^{G}\right)\right|}\right| = G. \quad (10)$$

Optimization Objective of Difficulty-Aware Group Policy Optimization (DGPO)

Difficulty-aware Question-level Weighting (DQW)

Core idea: prioritizes learning from more challenging yet solvable questions

Specifically, DQW assigns a weight λ_s to each question q_s as follows:

$$\lambda_s = B_{\mathrm{v}} \cdot \frac{\exp\left(D_s/T\right)}{\sum_{s=1}^{B_{\mathrm{v}}} \exp\left(D_s/T\right)}, \quad \text{where } D_s = -\operatorname{mean}\left(\{r_{si}\}_{i=1}^{G}\right). \tag{11}$$

Here, D_s is the negative mean reward across all responses of the question q_s, serving as a measure of its relative difficulty at the current training stage.

DGPO = DGAE + DQW

"balance-then-reweight"

Multi-Aspect Question Reformulation (MQR)

Core Instructions for Multi-Aspect Question Reformulation

1. **Background:** Add a story background that is not related to the core mathematical content of the given question, but seems to be related to the question. If the given question already has such a background, change it to a new, complexer background.
2. **Term:** Invent a new, abstract mathematical term to define a concept that is central to the given question, and restate the entire question using this term.
3. **Sub-Problem:** Convert a key numerical condition of the given question which have a definite value into an independent sub-problem. The sub-problem may belong to any branch of mathematics (e.g., algebra, geometry, number theory, combinatorics).

A critical constraint is that *all reformulations must preserve the original gold answer*.

The newly generated questions respectively challenge the policy model's ability to:

1. identify critical mathematical information amidst noise;
2. grasp abstract mathematical concepts; and
3. perform reasoning that requires multiple steps and cross-domain knowledge.

Experiments

高瓴人工智能学院
Gaoling School of Artificial Intelligence

Main Results

Table 1: Comparative results of methods trained on the MATH dataset using Qwen2.5-Math-7B.

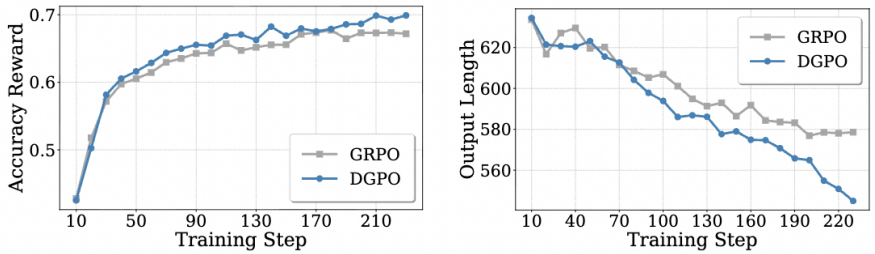
Methods	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg./Δ_{GRPO}
Base Model	12.19	4.79	35.23	48.60	15.07	16.33	22.04
GRPO	20.94	8.44	58.98	72.20	27.76	37.33	37.61
Dr.GRPO	21.04	8.23	58.59	72.05	28.58	35.89	37.40$_{-0.21}$
GPG	21.98	9.06	59.61	72.05	27.21	37.67	37.93$_{+0.32}$
DAPO	21.25	8.75	58.20	72.70	29.50	37.22	37.94$_{+0.33}$
GSPO	19.38	8.33	60.16	73.00	28.12	37.26	37.71$_{+0.10}$
GRPO-AD	21.56	9.48	59.06	73.25	29.14	37.07	38.26$_{+0.65}$
DGPO	23.85	10.21	**61.02**	74.25	31.07	38.33	39.79$_{+2.18}$
MQR	**25.00**	11.77	59.38	77.85	31.43	40.81	41.04$_{+3.43}$
MathForge	24.58	**12.60**	59.84	**79.95**	**33.36**	**42.67**	**42.17**$_{+4.56}$

Table 2: Comparative results of methods trained on the MATH dataset using varying base models.

Methods	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg./Δ_{GRPO}
Qwen2.5-Math-1.5B	6.87	3.65	30.94	34.95	8.55	21.93	17.82
+ GRPO	11.35	3.96	46.48	64.85	20.13	29.59	29.39
+ DGPO	11.25	5.73	49.84	65.45	21.14	30.85	30.71$_{+1.32}$
+ MQR	11.98	5.42	50.08	69.65	23.81	33.67	32.44$_{+3.05}$
+ MathForge	**13.23**	**7.71**	**52.34**	**70.10**	**25.74**	**33.89**	**33.84**$_{+4.45}$
Qwen2.5-3B	2.81	0.73	22.66	48.65	13.69	19.37	17.99
+ GRPO	5.31	1.56	33.28	63.35	22.89	26.41	25.47
+ DGPO	**6.98**	1.56	36.56	**65.80**	25.28	26.96	27.19$_{+1.72}$
+ MQR	5.10	1.56	39.53	65.20	25.74	29.19	27.72$_{+2.25}$
+ MathForge	5.73	**1.77**	**40.70**	65.40	**28.86**	**31.59**	**29.01**$_{+3.54}$
DeepSeek-Math-7B	0.42	0.10	13.28	31.05	9.56	9.00	10.57
+ GRPO	0.63	0.10	19.14	41.45	14.71	13.44	14.91
+ DGPO	1.98	0.42	21.02	41.85	18.93	15.00	16.53$_{+1.62}$
+ MQR	1.98	**0.83**	20.86	**44.25**	17.00	15.74	16.78$_{+1.87}$
+ MathForge	**3.12**	0.73	**21.72**	43.60	**20.68**	**16.74**	**17.77**$_{+2.86}$

Experiments

Analysis of DGPO



(a) Accuracy Reward. (b) Output Length.

Figure 1: Training dynamics of DGPO vs. GRPO evaluated on the MATH500 benchmark. Both models are trained on MATH using Qwen2.5-Math-7B.

Table 3: Ablation Results of DGPO trained on the MATH dataset using Qwen2.5-Math-7B.

Methods	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg./Δ_{GRPO}
GRPO	20.94	8.44	58.98	72.20	27.76	37.33	37.61
DGPO (w/o DGAE & DQW)	20.21	9.06	59.45	72.40	28.58	36.56	37.71$_{+0.10}$
DGPO (w/o DQW)	21.77	9.69	60.00	73.45	29.04	37.93	38.65$_{+1.04}$
DGPO (full)	**23.85**	**10.21**	**61.02**	**74.25**	**31.07**	**38.33**	**39.79**$_{+2.18}$
DGPO ($T = 1.0$)	23.12	9.06	59.45	74.15	30.61	37.78	39.03$_{+1.42}$
DGPO ($T = 2.0$)	**23.85**	10.21	61.02	74.25	**31.07**	**38.33**	**39.79**$_{+2.18}$
DGPO ($T = 5.0$)	22.81	**11.35**	60.55	73.80	30.42	38.26	39.53$_{+1.92}$
DGPO ($T = 10.0$)	21.35	9.79	**62.27**	**74.55**	29.96	37.67	39.27$_{+1.66}$

Table 4: Synergistic results of DGPO with other policy optimization methods trained on the MATH dataset using Qwen2.5-Math-7B.

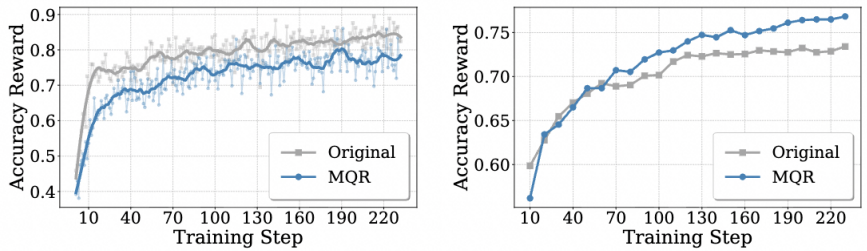
Methods	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Average
GPG	**21.98**	9.06	59.61	72.05	27.21	37.67	37.93
+ DGPO	21.77	**10.00**	**60.00**	**73.45**	**30.06**	**38.26**	**38.92**
DAPO	21.25	8.75	58.20	72.70	29.50	37.22	37.94
+ DGPO	**24.48**	**9.79**	**58.75**	**74.90**	**31.99**	**39.56**	**39.91**
GSPO	19.38	8.33	**60.16**	73.00	28.12	37.26	37.71
+ DGPO	**23.33**	**10.00**	59.14	**74.15**	**30.88**	**38.41**	**39.32**

Table 5: Comparative results of methods trained on the GEOQA-8k dataset using Qwen2.5-VL-3B-Instruct in the multimodal domain.

Methods	Base Model	GRPO	Dr.GRPO	GPG	DAPO	GSPO	GRPO-AD	DGPO
GeoQA/Δ_{GRPO}	39.79	57.43	57.96$_{+0.53}$	59.02$_{+1.59}$	59.02$_{+1.59}$	57.16$_{-0.27}$	58.09$_{+0.66}$	**59.95**$_{+2.52}$

Experiments

Analysis of DQW



(a) Training. (b) Evaluation.

Figure 2: Training dynamics of Original vs. MQR on training and evaluation data. Both models are trained on MATH and evaluated on MATH500 using Qwen2.5-Math-7B.

Table 6: Comparative results of methods trained on the original data vs. the MQR-augmented data using DGPO and varying base models.

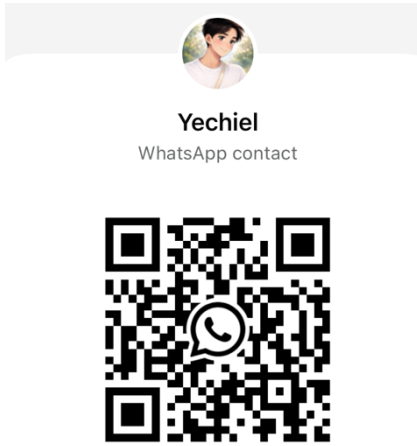
Models	Data	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Average
Qwen2.5-Math-7B	Ori.	**26.46**	9.17	58.67	74.65	31.62	38.81	39.90
	MQR	24.58	**12.60**	**59.84**	**79.95**	**33.36**	**42.67**	**42.17**
Qwen2.5-Math-1.5B	Ori.	11.98	5.21	50.62	68.40	24.26	32.59	32.18
	MQR	**13.23**	**7.71**	**52.34**	**70.10**	**25.74**	**33.89**	**33.84**
Qwen2.5-3B	Ori.	**6.04**	1.35	37.66	65.05	25.28	27.93	27.22
	MQR	5.73	**1.77**	**40.70**	**65.40**	**28.86**	**31.59**	**29.01**
DeepSeek-Math-7B	Ori.	2.19	0.21	21.02	**43.60**	18.29	14.52	16.64
	MQR	**3.12**	**0.73**	**21.72**	**43.60**	**20.68**	**16.74**	**17.77**

Table 7: Ablation Results of MQR on the MATH dataset using Qwen2.5-Math-7B.

Data	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg./$\Delta_{\text{Ori.}}$
Original	26.46	9.17	58.67	74.65	31.62	38.81	39.90
MetaMath-Rephrasing	25.21	11.35	59.45	76.70	31.71	39.93	40.73$_{+0.83}$
Original + Background	25.52	10.73	58.59	77.50	32.90	40.48	40.95$_{+1.05}$
Original + Term	25.52	11.15	58.98	77.75	33.09	40.93	41.24$_{+1.34}$
Original + Sub-Problem	**26.67**	10.94	58.75	77.05	**34.38**	41.36	41.53$_{+1.63}$
MQR	24.58	**12.60**	**59.84**	**79.95**	33.36	**42.67**	**42.17**$_{+2.27}$

Table 8: Comparative results of MQR using varying reformulator models on the MATH dataset.

Reformulators	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg./$\Delta_{\text{Ori.}}$
Original	26.46	9.17	58.67	74.65	31.62	38.81	39.90
Qwen2.5-7B-Instruct	25.10	11.98	58.67	76.85	33.00	40.96	41.09$_{+1.19}$
Qwen3-30B-A3B-Thinking	**25.73**	12.29	**59.84**	78.85	33.18	41.22	41.85$_{+1.95}$
OpenAI o3	24.58	**12.60**	**59.84**	**79.95**	**33.36**	**42.67**	**42.17**$_{+2.27}$



Thank you for your attention!