ICLR 2026

GPG: A Simple and Strong Reinforcement Learning Baseline for Model Reasoning

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, Yong Wang

AMAP, Alibaba Group

Project Page: https://github.com/AMAP-ML/GPG

Project Page

Motivation

	Components			
	Value Models	Reference Models	Surrogate Loss	Policy Constraint
PPO	✓	✓	✓	✓
GRPO	✗	✓	✓	✓
TRPO	✓	✗	✓	✓
GPG	✗	✗	✗	✗

- ***Reasoning gains increasingly rely on RL fine-tuning (RFT):*** *beyond SFT, RL encourages structured, multi-step reasoning and improves hard math/logic tasks.*
- ***Current RLHF/RFT methods are over-engineered for LLMs:*** *PPO/GRPO often require critics, reference models, surrogate losses, and KL control—making training complex and expensive.*
- ***LLM post-training is different from classic RL:*** *LLMs already have strong representations from pretraining, so many components designed for "learning from scratch" may be unnecessary.*
- ***Goal:*** *Develop a minimal, stable RL method that directly optimizes the original objective, while mitigating variance and bias—efficient and scalable.*

Method

Comparison of various RL methods

RL Method	Loss Function	Advantage Function
PPO (Schulman et al., 2017)	$\mathcal{L}_{\text{PPO}} = -\min\left[\frac{\pi_\theta(o)}{\pi_{\theta_{old}}(o)} \cdot A, \underbrace{\text{clip}\left(\frac{\pi_\theta(o)}{\pi_{\theta_{old}}(o)}, 1-\epsilon, 1+\epsilon\right)}_{\text{CLIP}} \cdot A\right]$	where A computed by applying GAE (Schulman et al., 2018) based on rewards and the **critic model**.
GRPO (Shao et al., 2024)	$\mathcal{L}_{\text{GRPO}} = -\left(\min\left[\frac{\pi_\theta(o)}{\pi_{\theta_{old}}(o)} \cdot A, \text{CLIP} \cdot A\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta \| \pi_{ref}\right]\right)$	$A = \frac{R(o) - \text{mean}\{R(o)\}}{\text{std}\{R(o)\}}$
Dr. GRPO (Liu et al., 2025a)	$\mathcal{L}_{\text{Dr.GRPO}} = \mathcal{L}_{\text{PPO}}$	$A = R(o) - \text{mean}\{R(o)\}$
DAPO (Yu et al., 2025)	$\mathcal{L}_{\text{DAPO}} = -\min\left[\frac{\pi_\theta(o)}{\pi_{\theta_{old}}(o)} \cdot A, \text{clip}\left(\frac{\pi_\theta(o)}{\pi_{\theta_{old}}(o)}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right) \cdot A\right]$	$A = \frac{R(o) - \text{mean}\{R(o)\}}{\text{std}\{R(o)\}}$
GPG	$\mathcal{L}_{\text{GPG}} = -\log \pi_\theta(o) \cdot A$	$A = \alpha * (R(o) - \text{mean}\{R(o)\})$

Core Objective of GPG

$$\mathcal{J}_{\text{GPG}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\left(-\log \pi_\theta(o_{i,t} \mid q, o_{i,<t})\hat{A}_{i,t}\right)\right], \hat{A}_{i,t} = \frac{r_i - \text{mean}(\{R_i\}_{i=1}^{G})}{F_{norm}}.$$

Accurate Gradient Estimation (AGE)

$B \longrightarrow$ Batch Size $\quad M \longrightarrow$ Number of Same-label (All-right/All-wrong) Samples

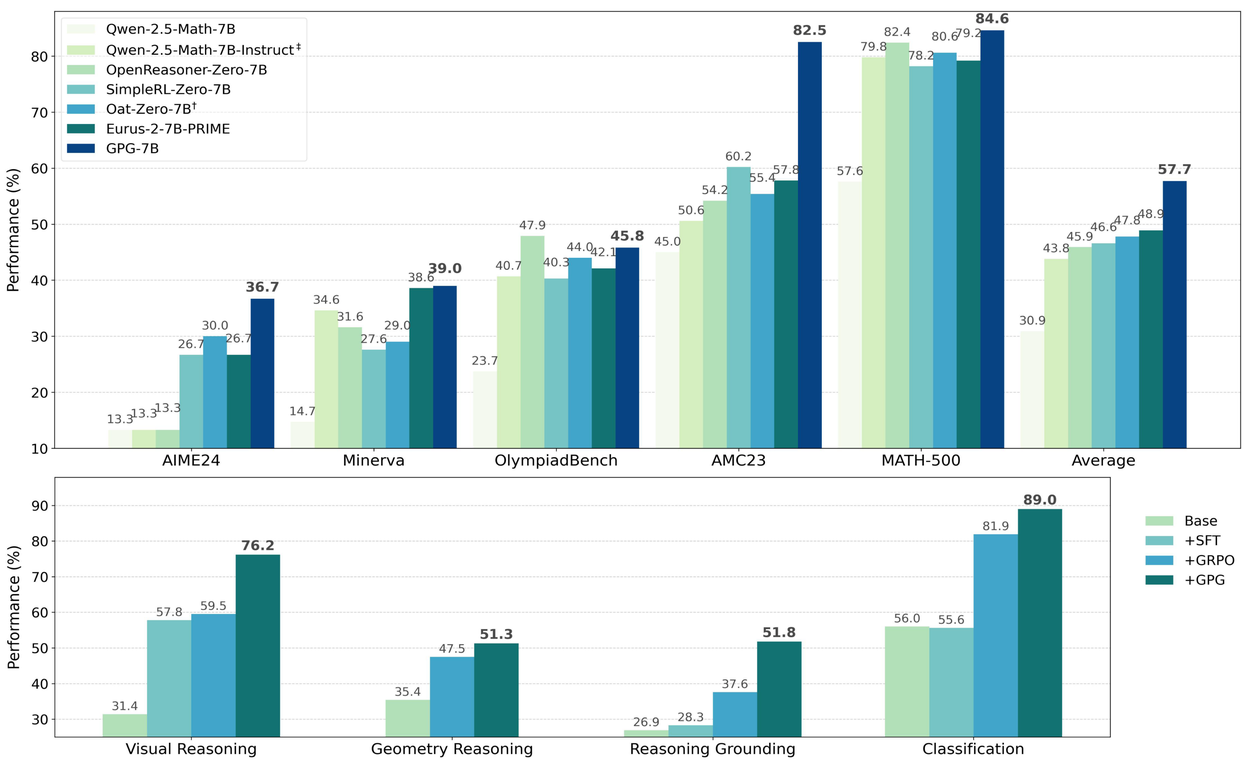
Standard Backpropagation (BP) Gradient

$$\mathbf{g} = \frac{\sum_{i=1}^{B}\mathbf{g_i}}{B} = \frac{\sum_{i=M+1}^{B}\mathbf{g_i}}{B}$$

AGE Gradient

$$\hat{\mathbf{g}} = \frac{\sum_{i=M+1}^{B}\mathbf{g_i}}{B-M} = \mathbf{g}\frac{B}{B-M} = \alpha\mathbf{g}, \alpha = \frac{B}{B-M}.$$

Experiment



Distilled 1.5B Models	Average	AIME24	MATH-500	AMC23	Minerva	OlympiadBench
DeepSeek-R1-Distill-Qwen-1.5B	48.9	28.8	82.8	62.9	26.5	43.3
Still-3-1.5B-Preview	51.6	32.5	84.4	66.7	29.0	45.4
Open-RS1†	53.1	33.3	83.8	67.5	29.8	50.9
Open-RS3†	52.0	26.7	85.4	70.0	27.9	50.2
GPG-RS1	55.7	33.3	87.6	77.5	29.4	50.5
GPG-RS3	55.5	33.3	85.0	80.0	26.8	52.4

7B Models	Average	AIME24	MATH-500	AMC23	Minerva	OlympiadBench
Qwen-2.5-Math-7B-Instruct ‡	43.8	13.3	79.8	50.6	34.6	40.7
Qwen2.5-Math-7B	30.9	13.3	57.6	45.0	14.7	23.7
Qwen2.5-Math-7B (no template)*	38.2	0.2	69.0	45.8	21.3	34.7
rStar-Math-7B (Guan et al., 2025)	-	26.7	78.4	47.5	-	47.1
Eurus-2-7B-PRIME (Cui et al., 2025)	48.9	26.7	79.2	57.8	38.6	42.1
Oat-Zero-7B (Liu et al., 2025a)	51.4	43.3	80.0	62.7	30.1	41.0
Oat-Zero-7B (Liu et al., 2025a)†	47.8	30.0	80.6	55.4	29.0	44.0
OpenReasoner-Zero-7B @ 8k (Hu et al., 2025)	45.9	13.3	82.4	54.2	31.6	47.9
SimpleRL-Zero-7B (Zeng et al., 2025)*	46.6	26.7	78.2	60.2	27.6	40.3
GPG-Zero-7B	57.7	36.7	84.6	82.5	39.0	45.8

Models	$\text{mIoU}_{\text{test}}$	mIoU_{val}	$\text{gIoU}_{\text{test}}$
Qwen2-VL-2B	26.9	30.1	25.3
+ SFT	28.3	29.7	25.3
+ GRPO	37.6	34.4	34.4
+ GPG	51.8	51.3	50.4

Models	Average	Flower102	Pets37	FGVC	Cars196
Qwen2-VL-2B	56.0	54.8	66.4	45.9	56.8
+ SFT	55.6	58.5	55.5	67.9	40.5
+ GRPO	81.9	71.4	86.1	74.8	95.3
+ GPG	89.0	79.3	90.8	88.5	97.5

Models	Total	Count	Relation	Depth	Distance
Qwen2-VL-2B	31.38	54.69	22.46	0.16	31.66
+ SFT	57.84	60.02	68.92	55.00	45.83
+ GRPO	59.47	59.64	66.76	54.16	56.66
+ GPG	76.15	66.62	83.23	81.66	75.50

Models	$\text{GEOQA}_{\text{Test}}$
Qwen2.5-VL-3B-Instruct	35.41
+ GRPO	47.48
+ GPG	51.33