Group Adaptive Clipping Policy Optimization

Anonymous ACL submission

Abstract

Group-relative policy optimization methods for reinforcement learning with verifiable rewards (RLVR) typically use a fixed importance-sampling (IS) ratio clipping boundary across all rollouts. We identify a key limitation of this design: rare correct rollouts on harder problems and abundant correct rollouts on easier problems are clipped at comparable rates, despite contributing very different learning signals. In particular, rollouts with low group success, few correct solutions within a rollout group, tend to exhibit larger IS ratios and carry stronger gradient signal for exploration and solving new problems, yet are disproportionately suppressed by uniform clipping.

To address this issue, we propose *Group Adaptive Clipping Policy Optimization (GAPO)*, a simple plug-in modification to group-relative policy optimization that adapts the IS clipping boundary to the rollout advantage. GAPO is motivated by a reverse-KL trust-region perspective, which suggests that rollouts with larger learning signal should receive proportionally greater update headroom. Importantly, GAPO requires no reward shaping or objective modification, preserving the standard PPO/GSPO surrogate while adapting only the clipping threshold. Across Qwen and Llama base models, GAPO consistently improves both pass@1 and pass@k over fixed clipping and advantage-shaping baselines on mathematical reasoning benchmarks.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) trains language models using PPO-style clipped policy optimization (Schulman et al., 2017), where a fixed importance-sampling (IS) ratio boundary approximates a trust region by suppressing updates once the policy moves too far from its previous iterate. In group-relative RLVR methods such as GRPO (Shao et al., 2024), GSPO (Zheng et al., 2025), and DAPO (Yu et al., 2026), rollouts are assigned advantages relative to other samples from the same prompt. Specifically, in a group of k rollouts with c correct solutions, correct rollouts receive advantage $A_i = (k - c)/k$, creating a structured spectrum ranging from scarce, high-advantage correct rollouts (low c) to abundant, low-advantage correct rollouts (high c).

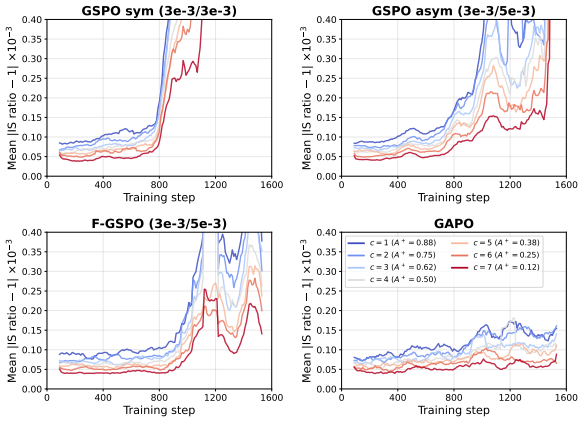

Figure 1: **Mean deviation of the importance-sampling (IS) ratio over RLVR training, binned by correctness count c within a group.** Two patterns hold across algorithms: (1) the IS ratio is consistently higher for scarcely correct rollouts (small c, fewer correct solutions per group), and (2) for these rollouts the ratio continues to grow at a rate equal to or exceeding that of rollouts on easier problems (large c). Together these patterns motivate adapting the clip boundary to question difficulty: under a fixed boundary, the rare high-advantage rollouts on hard problems hit the clip first and are suppressed disproportionately early.

A fixed clipping boundary treats this spectrum uniformly, despite substantial differences in learning signal. We observe that before clipping activates, IS ratios grow proportionally to advantage: scarce correct rollouts on harder problems move away from the reference policy substantially faster than abundant correct rollouts on easier problems (Figure 1). Yet once clipping begins, both

are truncated at comparable rates under a uniform boundary. This creates a mismatch: scarce correct rollouts, which provide valuable signal for exploration and solving new problems, are clipped too aggressively, while abundant redundant rollouts are granted equal update headroom despite contributing less to improvement.

A natural response might be to widen the clipping boundary uniformly, as explored in DAPO (Yu et al., 2026). However, this fails to address the core issue: the problem is not the absolute width of the clipping boundary, but its uniformity across rollouts with fundamentally different advantages. What must change is the *relative* clip width across different levels of group success.

Under a reverse-KL trust-region perspective, the optimal policy update at a single prompt allocates probability to responses proportionally to their advantage, yielding a trust-region-optimal importance-sampling (IS) ratio that scales exponentially with advantage (Chen et al., 2018). This suggests a simple principle for clipping: rollouts with larger learning signal should receive proportionally greater update headroom. Pushing an IS ratio substantially beyond this optimum either violates the trust region or over-allocates probability mass to a single response at the expense of alternatives, making the optimal ratio a natural guide for adaptive clipping.

We leverage this observation to introduce *Group Adaptive Policy Optimization (GAPO)*, a simple plug-in modification to group-relative policy optimization that adapts the per-rollout clip threshold to rollout advantage. In RLVR, binary rewards induce only a small number of discrete positive advantage levels, allowing the adaptive clipping threshold to be computed directly from the group success statistic c. GAPO preserves the standard PPO/GSPO surrogate objective and modifies only the clipping boundary. Unlike reward- or advantage-shaping approaches, GAPO continues to optimize pass@1 directly and does not exhibit the pass@1 drift that can arise from shaping objectives based on difficulty (Plyusov et al., 2026), inference-time pass@k (Walder and Karkhanis, 2026; Chen et al., 2025b; Tang et al., 2025), or diversity (Li et al., 2025). We make the following contributions:

- We identify a *clipping asymmetry* in RLVR: under uniform clip boundaries, scarce correct rollouts (low c, high advantage) are clipped at rates comparable to abundant correct rollouts, despite carrying substantially stronger learning signal (Figure 3). We show that a reverse-KL trust-region perspective (Equation 5, 6) naturally motivates *advantage-dependent clipping*, yielding a trust-region-optimal IS ratio that scales exponentially with advantage and reduces to a practical closed-form clipping rule in RLVR through the group statistic c (Equation 8).

- We introduce Group Adaptive Clipping Policy Optimization (GAPO), a simple plug-in modification to PPO/GSPO that adapts only the clipping boundary while preserving the standard surrogate objective and direct optimization of pass@1 (Equation 11).

- We demonstrate consistent improvements in pass@1 and pass@k across Qwen2.5-Math-1.5B, Llama-3.2-3B-Instruct, and DeepSeek-R1-Distill-Qwen-1.5B on mathematical reasoning benchmarks (Table 2, 1, Figure 4), while maintaining strong correlation between IS ratio and advantage throughout training (Figure 5).

2 Preliminaries

2.1 RL with Verifiable Rewards

We consider reinforcement learning with verifiable rewards (RLVR) for language model post-training. Given a prompt x drawn from a dataset $\mathcal{D}$, the policy π_θ generates a complete response $y \sim \pi_\theta(\cdot \mid x)$, which receives a binary verifiable reward $R(x, y) \in \{0, 1\}$ from an automated verifier. The training objective is the expected reward

$$J(\theta) = \mathbb{E}[R(x, y)], \quad x \sim \mathcal{D},\ y \sim \pi_\theta(\cdot \mid x). \tag{1}$$

In group-relative policy optimization with k rollouts per prompt $\{y_i\}_{i=1}^k \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$, a correct rollout i in a group with c correct has advantage

$$A_i = r_i - \bar{r} = 1 - \frac{c}{k} = \frac{k-c}{k}, \tag{2}$$

where $r_i \in \{0, 1\}$. Scarce correct rollouts (c=1) receive large advantage $(k-1)/k$, while abundant correct rollouts (c=k−1) receive small advantage $1/k$. In this work, we do not perform advantage normalization for unbiased advantage estimate (Liu et al., 2025).

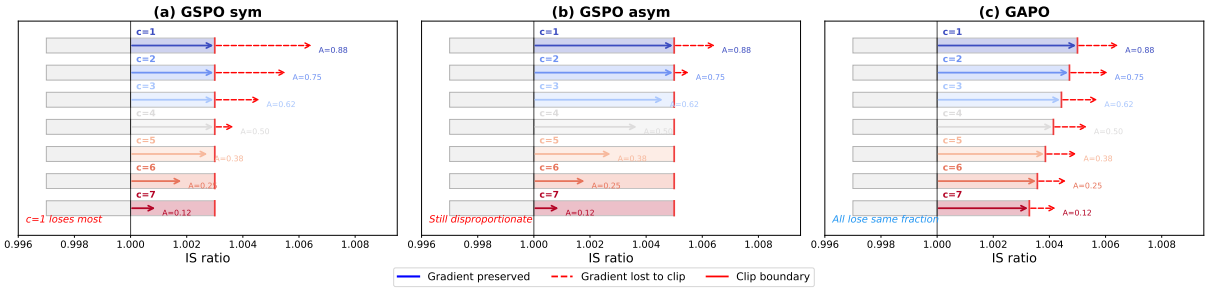

Figure 2: **(a) Symmetric clipping (b) Asymmetric Clipping (c) Adaptive Clipping.** Each row represents a correct rollout with group statistic c and advantage A_i. Solid arrows show preserved gradient; dashed red arrows show gradient lost to IS surpassing clipping thresholds. (a) GSPO symmetric: a uniform boundary clips scarce correct rollouts disproportionately. (b) GSPO asymmetric: wider upper boundary but still uniform across c. (c) GAPO: per-c adaptive boundary equation 11 keeps gradient updates proportional to advantage equation 8 even as clipping starts firing.

Standard PPO-style clipping applies a fixed trust region $[1-\epsilon,\ 1+\epsilon]$ uniformly to all rollouts. We observe that this disproportionately suppresses high-advantage (scarce) rollouts, which push their importance ratio past the clip boundary fastest.

2.2 Trust Region Policy Optimization

Trust regions and reverse KL. We build on the policy-improvement guarantee of Schulman et al. (2015). In the RLVR setting the bound takes the form

$$J(\theta) \geq L_{\pi_{\theta_{\text{old}}}}(\pi_\theta) - C \cdot \max_x D_{\text{KL}}(\pi_{\theta_{\text{old}}}(\cdot \mid x) \,\|\, \pi_\theta(\cdot \mid x)), \quad (3)$$

where the surrogate, written in importance-sampling form, is

$$\begin{aligned} L_{\pi_{\theta_{\text{old}}}}(\pi_\theta) &= \widehat{\mathbb{E}}_i\left[\frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)} A_i\right] \\ &= \widehat{\mathbb{E}}_i[\rho_i A_i], \end{aligned} \quad (4)$$

with $\widehat{\mathbb{E}}_i[\cdot]$ denoting the empirical expectation over a batch of rollouts $\{(x, y_i)\}$ drawn from $\pi_{\theta_{\text{old}}}(\cdot \mid x)$, and C depending only on the reward range. The bound follows from Pinsker's inequality applied to the worst-case total-variation distance between policies; maximising its right-hand side guarantees monotonic improvement of $J(\theta)$ at each update.

Equation 3 is stated with the forward direction $D_{\text{KL}}(\pi_{\theta_{\text{old}}} \,\|\, \pi_\theta)$, but the same guarantee holds with the reverse direction (Chen et al., 2018). The reason is symmetry: because $D_{\text{TV}}(P, Q) = D_{\text{TV}}(Q, P)$, Pinsker's inequality bounds the same quantity from either side, so $D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\theta_{\text{old}}})$ may replace $D_{\text{KL}}(\pi_{\theta_{\text{old}}} \,\|\, \pi_\theta)$ in equation 3 without weakening the guarantee. The constrained optimization we solve is therefore

$$\begin{aligned} &\max_\theta \ \widehat{\mathbb{E}}_i[\rho_i A_i] \\ &\text{s.t.} \ \widehat{\mathbb{E}}_i[D_{\text{KL}}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\theta_{\text{old}}}(\cdot \mid x))] \ \leq \ \delta, \end{aligned} \quad (5)$$

where the empirical expectation over the batch replaces the worst-case maximum for tractability, as in PPO (Schulman et al., 2017).

3 Group Adaptive Policy Optimization

3.1 Per-Prompt Trust Region Optimization

Following Chen et al. (2018), we consider local policy optimization at a state s, which is a prompt x in our case. The Lagrangian of maximizing equation 5 given x is

$$\begin{aligned} \mathcal{L}_x = &\sum_y \pi_\theta(y|x)\, A(x, y) \\ &- \lambda D_{\text{KL}}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\theta_{\text{old}}}(\cdot \mid x)). \end{aligned} \quad (6)$$

By solving the following Euler–Lagrange equation (Gelfand et al., 2000),

$$\begin{aligned} \frac{\partial \mathcal{L}_x}{\partial \pi_\theta(y|x)} &= A(x, y) - \lambda\left(\log \frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{old}}}(y|x)} + 1\right) \\ &= 0. \end{aligned} \quad (7)$$

we obtain the stationary point for the target policy $\pi^*_\theta(y|x) \ \propto \ \pi_{\theta_{\text{old}}}(y|x) \exp\left(\frac{A(x,y)}{\lambda}\right)$. At each rollout i for prompt x, the optimal IS ratio is

$$\rho^*_i = \frac{\pi^*_\theta(y_i|x)}{\pi_{\theta_{\text{old}}}(y_i|x)} \propto \exp(A_i/\lambda). \quad (8)$$

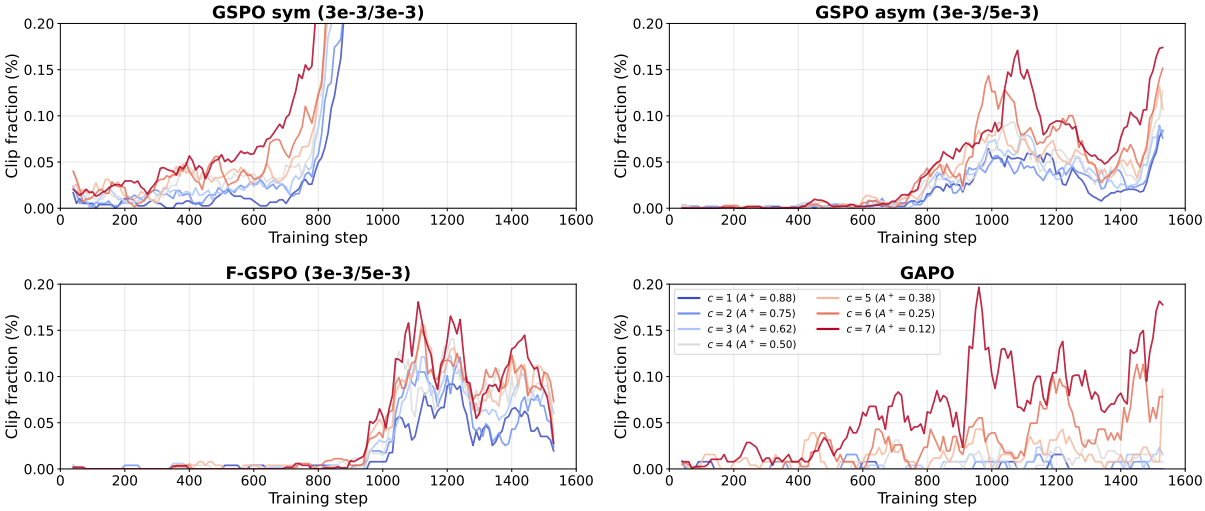

Figure 3: **Token clip fraction over training of Qwen2.5-Math-1.5B, by group correctness c.** Fraction of tokens clipped *within* rollouts at each c. Under fixed clipping (GSPO sym / asym), scarce correct rollouts on harder questions (low c, larger advantage A_i equation 2) are clipped at rates comparable to abundant correct rollouts (high c). GAPO clips high-c rollouts more and low-c rollouts less, retaining gradient on the rare correct rollouts most valuable for exploration.

Interpretation on the optimal IS ratio. ρ_i^* describes how much the probability of generating response y_i should be pushed up for prompt x to maximize the expected return on prompt x. Pushing ρ_i past ρ_i^* means either the trust region constraint is violated, or probability of generating some other response y_j is not pushed according to equation 8, lowering the expected return under this formulation. The clip should therefore fire at ρ_i^*, giving

$$\epsilon_{\text{hi}}(i) \;\geq\; \rho_i^* - 1 \;=\; \exp\left(\frac{A_i}{\lambda}\right) - 1. \qquad (9)$$

GSPO uses clip widths $\epsilon \sim 10^{-3}$, and GAPO inherits this scale, so $\epsilon_{\text{hi}} \sim 10^{-3}$. Inverting $\epsilon_{\text{hi}} = \exp(A_i/\lambda) - 1$ gives $|A_i/\lambda| = \log(1+\epsilon_{\text{hi}}) \approx 10^{-3}$, hence $\lambda \sim A_{\max}/10^{-3} \sim 10^3 \gg A_{\max}$ (with $A_{\max} = (k-1)/k \leq 1$). The linearization $\exp(A_i/\lambda) \approx 1 + A_i/\lambda$ then gives

$$\epsilon_{\text{hi}}(i) \;\approx\; \frac{A_i}{\lambda} \;=\; \frac{k - c_i}{k\lambda}. \qquad (10)$$

3.2 GAPO Adaptive Clip Formula

Normalising to interpolate between ϵ_{lo} (minimum, for $c = k$) and $\epsilon_{\text{hi}}^{\max}$ (maximum, for $c = 1$), we obtain the GAPO adaptive upper clip:

$$\boxed{\epsilon_{\text{hi}}(c) \;=\; \epsilon_{\text{lo}} + (\epsilon_{\text{hi}}^{\max} - \epsilon_{\text{lo}}) \cdot \frac{k - c}{k - 1}} \qquad (11)$$

with boundaries $\epsilon_{\text{hi}}(c{=}1) = \epsilon_{\text{hi}}^{\max}$, $\epsilon_{\text{hi}}(c{=}k) = \epsilon_{\text{lo}}$. Scarce correct rollouts (c=1) receive maximum headroom, while abundant correct rollouts ($c \to k$) are constrained to the minimum. Incorrect rollouts always use ϵ_{lo} (the upper clip is irrelevant for negative advantage).

Combined with GSPO's sequence-level importance ratio $s_i(\theta)$, the full GAPO objective is

$$\mathcal{L}^{\text{GAPO}} = \mathbb{E}\Big[\max\Big(-A_i \cdot s_i(\theta),$$
$$- A_i \cdot \text{clip}(s_i(\theta),\, 1 - \epsilon_{\text{lo}},\, 1 + \epsilon_{\text{hi}}(c_i))\Big)\Big], \qquad (12)$$

where the sequence-level ratio is the geometric mean of per-token ratios:

$$s_i(\theta) \;=\; \left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)}\right)^{1/|y_i|}. \qquad (13)$$

Both GAPO's adaptive clip and GSPO's sequence-level ratio operate at the rollout level, and the clip threshold directly gates the sequence-level ratio derived from the same per-rollout trust region argument.

4 Experiments & Results

We first compare the training dynamics of GAPO on Qwen2.5-Math-1.5B with three fixed clipping baselines. Specifically, we are interested in (i) how clip fraction within rollouts at each c differs between fixed and adaptive clipping (Figure 3), (ii) which clipping schedule maintains a high correlation between the empirical IS ratio and advantage

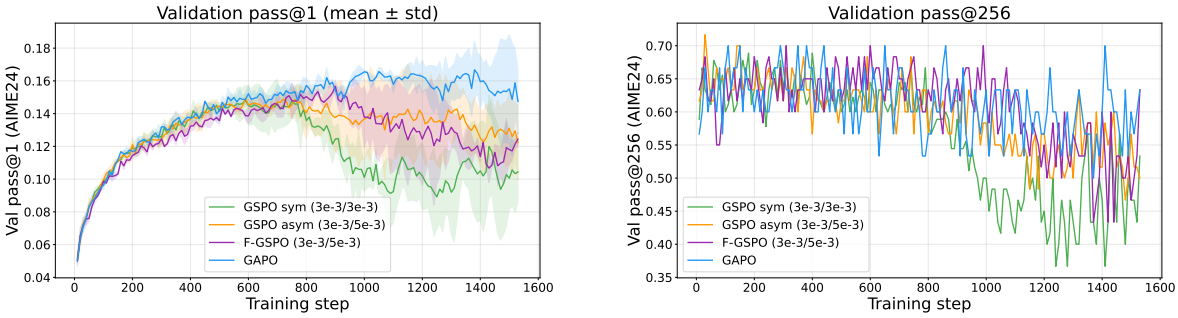

Figure 4: **Validation pass@1 and pass@256 on AIME24 over training.** Qwen2.5-Math-1.5B under four RLVR algorithms (mean ± std across seeds). GAPO sustains higher pass@1 and pass@256 in late training, while fixed-clip baselines (especially symmetric GSPO) show pass@256 collapse after step ~ 1000. The diversity loss adaptive clipping is designed to avoid by retaining gradient on low-c rollouts (Figures 3, 5).

(Figure 5), and (iii) whether this correlation translates into higher pass@1 and pass@k (Figure 4). We then benchmark GAPO against representative RLVR methods on two baes models without SFT (Qwen2.5-Math-1.5B, Llama-3.2-3B-Instruct) in Table 2 and one reasoning-distilled base model (DeepSeek-R1-Distill-Qwen-1.5B) in Table 1.

4.1 Experiment Setup

Models and datasets. We train and evaluate three base models spanning two model families and two pretraining regimes: Qwen2.5-Math-1.5B (math-pretrained) (Yang et al., 2025), Llama-3.2-3B-Instruct (general-purpose instruction-tuned) (Grattafiori et al., 2024), and DeepSeek-R1-Distill-Qwen-1.5B (reasoning-distilled). We include this post-SFT model, since their RLVR usually has a different training dynamic, and we aim to show GAPO works for all settings. All runs train on the DeepScaleR dataset with 39202 samples after filtering duplicates (Tan et al., 2026).

Training details. Our main method builds on GSPO's sequence-level importance ratio $s_i(\theta)$ equation 13, with clip boundaries $\epsilon_{\text{lo}} = 3\text{e}{-3}$ and $\epsilon_{\text{hi}}^{\max} = 5\text{e}{-3}$ set to the optimal values reported by Khatri et al. (2025). For token-IS, we use clip boundaries $\epsilon_{\text{lo}} = 3\text{e}{-3}$ and $\epsilon_{\text{hi}}^{\max} = 5\text{e}{-3}$. We use group size $k = 8$. Loss is aggregated at the token level (sum of per-token losses divided by total token count). We do not normalise the advantage by $\text{std}(\{R(q, o_1), \ldots, R(q, o_k)\})$, following Dr.GRPO (Liu et al., 2025); this std normalisation introduces a question-level difficulty bias, upweighting groups whose rewards are nearly all 1 or 0. Full hyperparameters are listed in Table 3.

Baselines. We compare against fixed-clip RLVR methods at matched clip widths: GRPO (Shao et al., 2024), Dr.GRPO (Liu et al., 2025), GSPO with symmetric and asymmetric clip ranges (Zheng et al., 2025), and the focal-shaping variants F-GRPO (Plyusov et al., 2026) and F-GSPO. These apply advantage shaping $\widetilde{A}_i = (1 - c/k)^{\gamma} A_i$ with $\gamma \geq 0$. This selection spans the standard group-relative baselines and the two main alternative directions for adapting the surrogate (advantage shaping and asymmetric clipping) for exploration. The symmetric clipping (ϵ_{lo}=$3e{-}3$, $\epsilon_{\text{hi}}^{\max}$=$3e{-}3$) baseline is also chosen to show that GAPO's improvement comes not from clipping tighter, but from the relative difference in clip width across rollouts with different group success rates. We also evaluate a token-IS variant of GAPO against token-IS baselines. Among non-adaptive methods, GSPO's sequence-level IS is the strongest baseline in our setting, so we adopt it as the base for both GAPO and the advantage-shaping baseline F-GSPO.

Evaluation. We evaluate on standard mathematical reasoning benchmarks (AIME24 (Gokce et al., 2024), AIME25 (Ye et al., 2025), AMC, MATH500 (Hendrycks et al., 2021), Minerva (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024)) and report IFEval (Zhou et al., 2023) as an out-of-domain probe for instruction-following retention. For Qwen2.5-Math-1.5B (Yang et al., 2025) and Llama-3.2-3B-Instruct (Grattafiori et al., 2024), we report pass@1 and pass@256 with temperature 1.0, top-p 1, $n = 256$ samples per prompt, and $T_{\max}$ of 3092 and 8192 respectively. For DeepSeek-R1-Distill-Qwen-1.5B we report pass@1 and pass@16 with temperature 0.6, top-p 1, $n = 16$, and $T_{\max} = 24576$ to accommodate longer rollouts.

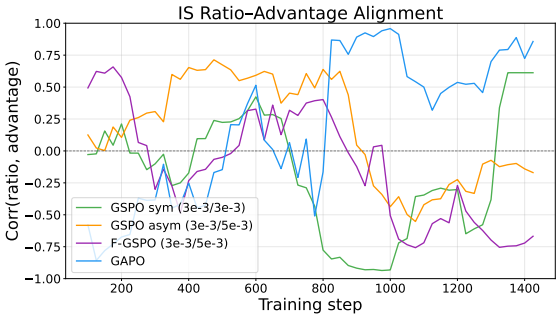

Figure 5: **IS–advantage correlation over training.** Windowed Pearson correlation between per-c IS-ratio deviation and advantage ($W = 200$ steps). Positive correlation reflects the trust-region-optimal relationship $\rho_i^* \propto \exp(A_i/\lambda)$ equation 8. GAPO maintains $r > 0.8$ in late training; fixed-clip baselines degrade toward negative correlation.

4.2 Adaptive clipping preserves the IS–advantage correlation

We compare GAPO against three fixed-clip baselines on Qwen2.5-Math-1.5B across clip fraction (Fig. 3), IS–advantage correlation (Fig. 5), and validation pass@1 / pass@256 (Fig. 4).

Under the reverse-KL trust region at a single prompt, the optimal target IS ratio equation 8 is proportional to advantage (to first order). We do not directly optimise toward this target; we instead clip proportional to advantage. Even so, Figure 5 shows GAPO retains a high correlation between the importance-sampling ratio and advantage throughout training, whereas fixed clipping breaks the correlation by clipping scarce correct rollouts at the same fraction as abundant correct ones.

Step0 – Step400. Qwen2.5-Math-1.5B has no reasoning capability, so the correlation at step 0 reflects sampling noise from RL and datasampler initialization rather than RL dynamics.

Step400 – Step600. This is a clip-free regime for scarce correct rollouts from low c groups. This results in a consistent rise in correlation across all baseline methods. Figure 3 shows clipping fires infrequently in this window, so the advantage–ratio coupling is preserved.

Step600+. The baselines only diverge once the clip fraction rises, at which point GAPO sustains the correlation while fixed clipping collapses it. Figure 4 confirms that GAPO's preserved IS–advantage structure improves pass@1 while retaining high pass@256. Beyond aggregated metrics, Figure 7 shows GAPO maintains problem coverage. While baselines progressively lose solve rates on medium-difficulty AIME24 problems, GAPO retains 10.9 problems with more than 5% solve rate in late training compared to 6.7-8.6 for baselines. Appendix B shows an example of diverse reasoning paths.

Base model: DeepSeek-R1-Distill-Qwen-1.5B Pass@1 / Pass@16 (Temperature=0.6, Top-p=1, n=16, Tmax=24576)				
In-domain	**Base**	**GSPO**	**F-GSPO**	**GAPO**
AIME24	28.5/60.0	41.3/73.3	40.2/73.3	**44.0/76.7**
AIME25	22.3/43.3	29.4/50.0	29.6/46.7	**30.8/56.7**
AMC	71.4/**95.0**	79.2/**95.0**	82.2/**95.0**	**83.44/95.0**
MATH500	70.4/82.4	83.0/**94.7**	**85.5**/93.2	**85.5**/93.2
Minerva	19.0/47.4	26.6/50.4	27.9/44.3	**29.1**/48.9
Olympiad	41.5/60.3	57.8/76.2	59.0/72.4	**59.4/76.2**

Table 1: **Comparison of GAPO with GSPO baselines on DeepSeek-R1-Distill-Qwen-1.5B.** Pass@1 / Pass@16 across six in-domain math benchmarks at group size N=8. *Base* is the pretrained model evaluated without RL training.

4.3 Benchmark results

Tables 1 and 2 report pass@1 / pass@k across three base models and the seven benchmarks listed in the setup. GAPO is compared against fixed-clip RLVR methods (GRPO, Dr.GRPO, symmetric and asymmetric GSPO) and the focal-shaping variants (F-GRPO, F-GSPO) at matched clip widths.

DeepSeek-R1-Distill-Qwen-1.5B. On a reasoning-distilled base model with longer rollouts ($T_{\max} = 24576$), GAPO leads pass@1 on all six in-domain benchmarks (Table 1), with the largest gains on AIME24 ($44.0/76.7$, $+2.7/+3.4$ over GSPO) and AIME25 ($30.8/56.7$, $+1.4/+6.7$ over GSPO).

Qwen2.5-1.5B-Math. GAPO achieves the best average pass@1 / pass@256 ($37.9/76.3$) among GSPO variants, and the strongest AIME24 pass@1 (17.9, vs. 15.4 for F-GSPO and 16.2 for asymmetric GSPO). At the token-IS clip range ($\epsilon_{\text{lo}} = 0.2$), GAPO-token-IS outperforms GRPO, F-GRPO, and Dr.GRPO on average and leads pass@1 on five of the six math benchmarks, with the largest gains on AIME24 pass@1 (14.6) and IFEval (20.9 vs. 11.4–19.4 for the baselines).

Llama-3.2-3B-Instruct. GAPO again leads on average pass@1 / pass@256 ($24.1/63.4$) and substantially improves AIME24 pass@1 over the fixed-clip GSPO variants (14.6 vs. 10.7–14.4). GAPO also achieves the best IFEval (56.4), tying F-GRPO.

Method	ϵ_{lo}	ϵ_{hi}	In-domain							OOD
			Avg.	AIME24	AIME25	AMC	MATH500	Minerva	Olympiad	IFEval
Qwen2.5-1.5B-Math			**Pass@1 / Pass@256 (Temperature=1.0, Top-p=1, n=256, Tmax=3092)**							
GRPO	0.2	0.2	36.7/74.4	13.8/61.1	9.9/58.0	53.1/96.2	75.4/95.6	31.9/61.1	36.3/74.3	12.2
F-GRPO	0.2	0.2	36.3/74.5	13.0/60.7	10.5/57.9	51.6/95.9	74.7/96.1	31.0/61.0	37.0/75.5	11.4
DrGRPO	0.2	0.28	33.8/75.1	13.1/60.0	8.85/63.3	51.8/97.5	70.5/94.6	22.3/62.1	36.0/73.2	19.4
GAPO-token-IS	0.2	0.28	**37.6/75.6**	**14.6/63.3**	10.2/60.0	**54.9/97.5**	**75.8/96.4**	**32.3/62.1**	**37.6**/74.2	**20.9**
GSPO	3e−3	3e−3	36.8/76.2	15.3/60.0	9.28/60.0	53.0/100	75.2/97.0	30.9/62.9	37.3/77.3	22.1
GSPO	3e−3	5e−3	37.7/74.1	16.2/60.0	9.54/56.7	55.3/95.0	76.3/95.6	31.0/63.6	37.6/73.7	20.1
F-GSPO	3e−3	5e−3	36.9/76.1	15.4/63.3	9.44/56.7	53.7/97.5	75.0/97.0	30.7/63.4	37.1/78.5	21.0
GAPO	3e−3	5e−3	**37.9/76.3**	**17.9/63.3**	**10.6**/56.7	53.4/97.5	**76.5/97.5**	**31.2/64.1**	**38.0/78.5**	**22.2**
Llama3.2-3B-Instruct			**Pass@1 / Pass@256 (Temperature=1.0, Top-p=1, n=256, Tmax=8192)**							
GRPO	0.2	0.2	23.0/59.9	10.7/40.7	0.7/21.5	30.5/88.2	55.0/90.6	21.8/59.0	19.4/59.3	54.1
F-GRPO	0.2	0.2	23.0/63.4	12.1/46.1	1.0/29.5	29.8/90.6	54.1/92.9	21.0/60.1	20.1/61.3	56.4
GSPO	3e−3	3e−3	22.7/56.5	13.6/36.7	0.4/16.7	27.8/86.5	51.5/87.8	20.3/57.6	22.3/53.9	54.9
GSPO	3e−3	5e−3	23.0/59.2	14.4/46.7	0.7/20.0	27.9/87.5	52.0/89.2	20.5/55.7	22.4/56.1	54.8
F-GSPO	3e−3	5e−3	23.8/59.2	13.2/36.7	0.9/23.3	31.1/92.5	54.0/87.8	21.1/57.6	22.5/57.5	56.3
GAPO	3e−3	5e−3	**24.1/63.4**	**14.6/53.3**	**1.0**/23.3	**31.6**/91.2	54.6/**93.2**	20.3/**60.5**	22.4/58.9	**56.4**

Table 2: **Comparison of GAPO with fixed-clip RLVR baselines on two base models.** Pass@1 / Pass@256 on six in-domain math benchmarks (Avg., AIME24, AIME25, AMC, MATH500, Minerva, Olympiad) and one OOD benchmark (IFEval) for Qwen2.5-1.5B-Math and Llama-3.2-3B-Instruct, trained at group size N=8. Columns ϵ_{lo} and ϵ_{hi} give the lower and upper clip thresholds (equal values denote symmetric clipping; differing values denote asymmetric clipping). Blue shaded rows mark our methods. GRPO and F-GRPO results are taken from Plyusov et al. (2026).

Across all three base models, GAPO's advantage-proportional clipping yields consistent pass@1 improvements without sacrificing pass@k diversity, supporting the claim that GAPO redistributes gradient toward scarce correct rollouts without overfitting on abundant ones.

4.4 Ablation Study

We ablate two key algorithm choices for GAPO. First, in Table 2, the more principled sequence-level adaptive IS clipping (AIME24 pass@1 17.9%, math average pass@1 37.9%) outperforms token-level adaptive IS clipping (AIME24 pass@1 14.6%, math average pass@1 37.6%). Second, the number of rollouts per group k is the critical hyperparameter: increasing k from 4 to 8 widens the gap between adaptive (GAPO) and fixed (GSPO) clipping (Figure 6). With GAPO, increasing the number of rollouts per group k benefits both the advantage estimate and the resolution of the per-prompt clip threshold. On this training dataset, further increases beyond k=8 yield no additional gain, so we conduct all other training with k=8 for efficiency.

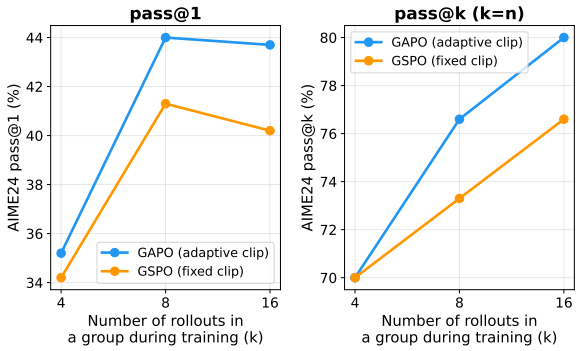

Figure 6: Ablation on the number of rollouts per group k used during training. Scaling from k=4 to k=8 yields clear improvement in both pass@1 and pass@256, and widens the gap between fixed clipping (GSPO) and adaptive clipping (GAPO).

5 Related Work

5.1 LLM Post-training

SFT. s1 (Muennighoff et al., 2025) shows that SFT on only 1K high-quality DeepSeek-R1 (Guo et al., 2025) traces can surpass o1-preview (Jaech et al., 2024). SSFT (Jia et al., 2025) uses a matching loss to distill from multiple teachers to capture diverse reasoning modes. On-policy distillation (Agarwal et al., 2024) leverages the access to teacher logits to perform token-level distribution matching between

the student and the teacher.

RLVR. Khatri et al. (2025) surveys group-relative RLVR methods. For token loss aggregation, Dr.GRPO (Liu et al., 2025) normalizes by a constant and DAPO (Yu et al., 2026) by total batch tokens, both to avoid length bias. For advantage normalization, Dr.GRPO removes the within-group std rescaling to preserve unbiasedness, while REINFORCE++ (Hu et al., 2025a) substitutes batch-level std. All use token-level IS with fixed clipping, with DAPO widening the upper threshold asymmetrically. GSPO (Zheng et al., 2025) instead uses sequence-level IS and clipping. CISPO (Chen et al., 2025a) clips the IS weight directly while preserving every token's gradient, a weighted REINFORCE. DAPO's clip-higher and CISPO's gradient preservation share our motivation of retaining exploration tokens. We arrive at the same goal through an explicit closed-form derivation: the per-prompt reverse-KL trust-region-optimal IS ratio suggests a clip threshold proportional to advantage. Other exploration-oriented approaches scale the number of rollouts (Hu et al., 2025b), introduce separate fixed clipping thresholds for correct and incorrect rollouts (Karaman et al., 2026), or only optimize policy on exploratory forking tokens (Wang et al., 2025b,a).

5.2 Advantage and Reward Shaping for Exploration in RLVR

Yue et al. (2025) found that pass@k of RLVR-trained LLMs score below the base model at large k, motivating a line of work that directly optimizes pass@k (Chen et al., 2025b; Peng et al., 2025; Walder and Karkhanis, 2026; Thrampoulidis et al., 2025). However, pass@1, the deployment metric, has been observed to degrade under this training (Walder and Karkhanis, 2026; Barakat et al., 2026). Rather than directly optimizing pass@k, others propose dedicated advantage-shaping schemes to encourage exploration. F-GRPO (Plyusov et al., 2026) and He et al. (2025) apply advantage shaping based on within-group success count; Zhou et al. (2025) reweights the loss across groups by success rate; and Gai et al. (2025) adds a differential-smoothing reward based on per-rollout correctness. In contrast, GAPO still optimizes the pass@1 objective, encouraging exploration by maintaining high correlation between the empirical IS ratio and advantage, which preserves gradients on rare correct rollouts relative to redundant ones, according to the target IS ratio for the reverse KL trust region formulation.

6 Conclusion

We provide a simple plug-in adaptive clipping method, Group Adaptive Policy Optimization (GAPO), motivated by the reverse-KL trust-region-optimal IS ratio at a single prompt. Under this trust region, the optimal IS ratio scales exponentially with advantage, suggesting that the per-rollout clip threshold should scale with advantage rather than be applied uniformly. In RLVR, binary rewards reduce this to a closed-form schedule indexed by the group correctness count c, requiring no new hyperparameters beyond the existing clip range. Across Qwen2.5-Math-1.5B, Llama-3.2-3B-Instruct, and DeepSeek-R1-Distill-Qwen-1.5B on mathematical reasoning benchmarks, GAPO consistently improves pass@1 while retaining pass@k, and maintains a high IS–advantage correlation throughout training where fixed clipping collapses it. Unlike reward-shaping and advantage-shaping approaches for exploration, GAPO still optimizes the deployment metric pass@1 directly while implicitly encouraging exploration through optimally allocating trust-region constraint.

Limitations

Joint adaptive clipping. We adapt only the upper clip threshold ϵ_{hi}, which governs positive-advantage rollouts. The lower clip ϵ_{lo} for negative-advantage (incorrect) rollouts is held fixed throughout. Adapting ϵ_{lo} analogously, and the joint adaptation of $(\epsilon_{\text{lo}}, \epsilon_{\text{hi}})$, is left to future work. **Multi-prompt batch analysis.** Our derivation of the per-rollout optimal IS ratio ρ_i^* is based on the reverse-KL trust region at a single prompt x (Section 3.1); extending this analysis to the multi-prompt training setting is left open. The empirical IS–advantage correlations we observe (Figure 5) provide evidence that the per-prompt heuristic is not vacuous as a guide for adaptive clipping in practice.

Ethical Considerations

Scope and broader impact. GAPO is a methodological contribution to RLVR: it adapts the per-rollout clipping threshold in policy optimization without introducing new model capabilities, training data, or evaluation protocols beyond those standard in the field. The intended impact is to improve the training efficiency and final performance

of LLMs on mathematical reasoning, with downstream applications in education, scientific computing, and assistive tools.

Data and models. All training data (DeepScaleR) and evaluation benchmarks (AIME24/25, MATH500, AMC, Olympiad, Minerva, IFEval) used in this work are publicly available. The base models we train (Qwen2.5-Math-1.5B, Llama-3.2-3B-Instruct, DeepSeek-R1-Distill-Qwen-1.5B) are openly released. No human subjects, annotators, or proprietary data were involved.

Dual-use considerations. As with any improvement to LLM training, GAPO may indirectly contribute to producing more capable reasoning systems, which carry the general dual-use considerations common to language modeling research. GAPO does not specifically lower the cost of generating harmful content, nor does it target capabilities relevant to harmful applications; we do not anticipate risks beyond those inherent to RLVR training in general.

Acknowledgements

All training and evaluation code will be released publicly upon acceptance.
We used Claude Opus 4.7 (Anthropic, 2026) in May 2026 for occasional language polishing only. This was done when we wanted to make sure there were no grammatical errors in a few sentences. No text, code, experiment results, or figures were generated by the LLM. We made our own hypotheses, completed all technical content on our own, and drew our own conclusions. We also verified all outputs ourselves. The occasional grammar checks were done by typing into the LLM chatbox.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pages 21246–21263.

Anthropic. 2026. Claude 4.7 Opus. https://claude.ai. Generative AI Chatbot.

Anas Barakat, Souradip Chakraborty, Khushbu Pahwa, and Amrit Singh Bedi. 2026. Why pass@ k optimization can degrade pass@ 1: Prompt interference in llm post-training. *arXiv preprint arXiv:2602.21189*.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, and 1 others. 2025a. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*.

Gang Chen, Yiming Peng, and Mengjie Zhang. 2018. An adaptive clipping approach for proximal policy optimization. *arXiv preprint arXiv:1804.06461*.

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. 2025b. Pass@ k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*.

Jingchu Gai, Guanning Zeng, Huaqing Zhang, and Aditi Raghunathan. 2025. Differential smoothing mitigates sharpening and improves llm reasoning. *arXiv preprint arXiv:2511.19942*.

Izrail Moiseevitch Gelfand, Richard A Silverman, and 1 others. 2000. *Calculus of variations*. Courier Corporation.

L. Gokce, A. Carratu, and S. Kanbir. 2024. *American Invitational Mathematics Examination (AIME) Preparation: 5 Practice Tests*. Math Topia Academy.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Andre Wang He, Daniel Fried, and Sean Welleck. 2025. Rewarding the unlikely: Lifting grpo beyond distribution sharpening. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25559–25571.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

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

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025a. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*.

Jian Hu, Mingjie Liu, Ximing Lu, Fang Wu, Zaid Harchaoui, Shizhe Diao, Yejin Choi, Pavlo Molchanov, Jun Yang, Jan Kautz, and 1 others. 2025b. Brorl: Scaling reinforcement learning via broadened exploration. *arXiv preprint arXiv:2510.01180*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Sheng Jia, Xiao Wang, and Shiva Prasad Kasiviswanathan. 2025. Training large language models to reason in parallel with global forking tokens. *arXiv preprint arXiv:2510.05132*.

Batuhan K Karaman, Aditya Rawal, Suhaila Shakiah, Mohammad Ghavamzadeh, Mingyi Hong, Arijit Biswas, and Ruida Zhou. 2026. Dispo: Enhancing training efficiency and stability in reinforcement learning for large language model mathematical reasoning. *arXiv preprint arXiv:2602.00983*.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S Dhillon, David Brandfonbrener, and Rishabh Agarwal. 2025. The art of scaling reinforcement learning compute for llms. *arXiv preprint arXiv:2510.13786*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. 2025. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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

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

Daniil Plyusov, Alexey Gorbatovski, Boris Shaposhnikov, Viacheslav Sinii, Alexey Malakhov, and Daniil Gavrilov. 2026. F-grpo: Don't let your policy learn the obvious and forget the rare. *arXiv preprint arXiv:2602.06717*.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Sijun Tan, Michael Luo, Justin Wong, Colin Cai, Xiaoxiang Shi, William Yuan Tang, Manan Roongta, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2026. Deepscaler: Effective RL scaling of reasoning models via iterative context lengthening.

Yunhao Tang, Kunhao Zheng, Gabriel Synnaeve, and Rémi Munos. 2025. Optimizing language models for inference time objectives using reinforcement learning. *arXiv preprint arXiv:2503.19595*.

Christos Thrampoulidis, Sadegh Mahdavi, and Wenlong Deng. 2025. Advantage shaping as surrogate reward maximization: Unifying pass@ k policy gradients. *arXiv preprint arXiv:2510.23049*.

Christian Walder and Deep Tejas Karkhanis. 2026. Pass@ k policy optimization: Solving harder reinforcement learning problems. *Advances in Neural Information Processing Systems*, 38:152416–152445.

Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. 2025a. Emergent hierarchical reasoning in llms through reinforcement learning. *arXiv preprint arXiv:2509.03646*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others.

2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yixin Ye, Yang Xiao, Tiantian Mi, and Pengfei Liu. 2025. Aime-preview: A rigorous and immediate evaluation framework for advanced mathematical reasoning. https://github.com/GAIR-NLP/AIME-Preview. GitHub repository.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2026. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Jingyu Zhou, Lu Ma, Hao Liang, Chengyu Shen, Bin Cui, and Wentao Zhang. 2025. Daro: Difficulty-aware reweighting policy optimization. *arXiv preprint arXiv:2510.09001*.

A Per-Problem Solve Rate Analysis

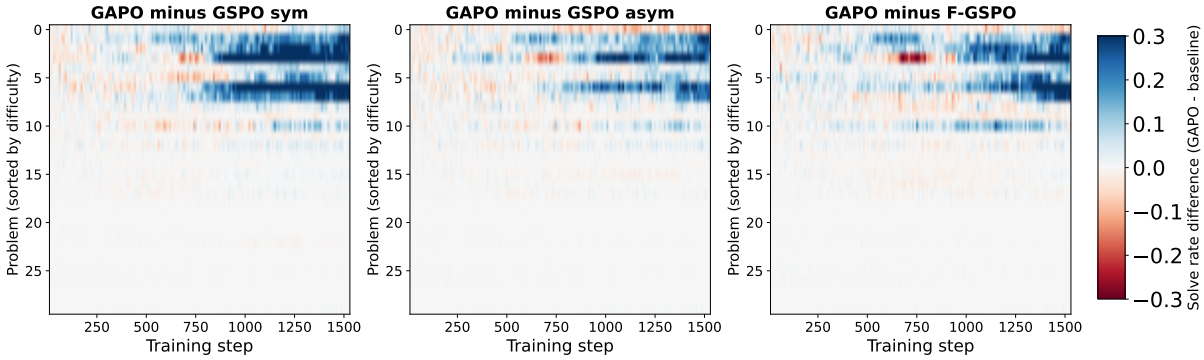

Figure 7: **Per-problem solve rate difference (GAPO minus baseline) on AIME24 validation over training for Qwen2.5-Math-1.5B** (256 samples per problem, evaluated every 10 steps). Problems are sorted by difficulty (easiest at top). Blue indicates higher solve rate for GAPO; red indicates higher solve rate for the baseline. Results are averaged across all seeds per method. GAPO's advantage is concentrated on medium-difficulty problems (rows 5–12) and grows in late training as baselines progressively lose solve rate on these frontier problems. Averaged over steps 1200–1530, GAPO maintains 10.9 problems with $> 5\%$ solve rate compared to 8.6 (F-GSPO), 8.5 (GSPO asym), and 6.7 (GSPO sym).

B Diverse Correct Solutions Found by GAPO

Problem (AIME24, Chip Placement)

There is a collection of 25 indistinguishable white chips and 25 indistinguishable black chips. Find the number of ways to place some of these chips in the 25 unit cells of a 5×5 grid such that:

(i) each cell contains at most one chip;

(ii) all chips in the same row and all chips in the same column have the same colour;

(iii) any additional chip placed on the grid would violate one or more of the previous two conditions.

`<Generation 1>` — Color the row/columns, add chips by deduction

`<|im_start|>assistant`
`<|im_start|>`
"each cell at most one chip, all chips in same row same color, same column same color" ... ... "rows and columns alternate in color" ... ... "mixing: choose which rows are white and which are black; similarly for columns" ... ... $2^5 - 2 = 30$ ways to choose rows (excluding all-white and all-black) ... ... $2^5 - 2 = 30$ ways for columns ... ... mixed configurations: $30 \times 30 = 900$ ... monochrome: 2 ... (*Comment:* each row/column picks white or black independently; cells where row and column colors agree get a chip, the rest stay empty. Excluding the trivial all-white and all-black choices gives $30 \times 30 = 900$ mixed placements; add the 2 monochrome grids: 902.)
`<|im_start|>answer`
Answer: ... $\boxed{902}$

<Generation 2> — Partition the grid into color blocks

<|im_start|>assistant
<|im_start|>
"the grid is divided into two rectangles of the same color" ... ... "all 25 cells same color: 2 configurations" ... ... "choose a subset of rows and a subset of columns to form the two regions" ... ... $2^5 - 2$ ways to divide the rows (any non-empty proper subset) ... $2^5 - 2$ ways to divide the columns ... $(2^5 - 2) \times (2^5 - 2) = 30 \times 30 = 900$ partitions ... total $2 + 900 = 902$ ... (*Comment:* each valid chip placement corresponds to a partition of the grid into sub-grids; counting non-trivial row/column splits gives 30×30, and the two trivial monochrome cases add 2.)
<|im_start|>answer
Answer: ... $\boxed{902}$

<Generation 3> — Alternating patterns

<|im_start|>assistant
<|im_start|>
"no two chips of different colors are adjacent" ... ... "each row and each column must be uniform in color" ... ... "the grid is bipartite in terms of color" ... ... explicit patterns enumerated: rows $\{1, 3, 5\}$ white, rows $\{2, 4\}$ black; or vice versa; ... "choose any subset of the 5 rows to be all white, excluding all-W and all-B: $2^5 - 2 = 30$" ... "similarly 30 choices for columns" ... $30 \times 30 = 900$ ... plus 2 monochrome ... (*Comment:* start from a specific pattern, e.g., rows $\{1, 3, 5\}$ white and rows $\{2, 4\}$ black, recognise that any way of splitting the 5 rows into white and black works as long as both colors appear, and likewise for the columns; this generalizes to 30×30 mixed patterns, plus the 2 single-color grids.))
<|im_start|>answer
Answer: ... $\boxed{902}$

Figure 8: **Diversity preservation on a single AIME24 problem.** The baseline GSPO produces only 3 correct generations on this problem, while GAPO produces 28 and, near peak validation pass@1, samples three distinct correct reasoning paths. All three converge to the same answer, but enter the problem from different conceptual angles. This suggests that GAPO retains multiple correct modes where the baseline has effectively collapsed.

C Training Hyperparameters

Parameter	Value
	Model
Base model	Qwen2.5-Math-1.5B / Llama3.2-3B-Instruct / DeepSeek-R1-Distill-Qwen-1.5B
Precision	bfloat16
	Optimization
Optimizer	AdamW (Loshchilov and Hutter, 2017)
Learning rate	1×10^{-6}
LR schedule	Constant
Weight decay	0.01
Gradient clip	1.0
Batch size (prompts)	256
Training steps	10 Epochs
	RL Settings
Rollouts per prompt k	8
KL coefficient λ	0
Entropy coefficient	0
Clip range $(\epsilon_{\text{lo}}, \epsilon_{\text{hi}})$	$(0.2, 0.28)$ for token IS, $(3e-3, 5e-3)$ for sequence IS
Reward	Verifiable (0/1)
	Generation
Training Dataset	DeepScaleR (Tan et al., 2026) with 39,202 samples.
Max prompt length	2048
Max response length	3,092 for Qwen2.5-Math / 8192 Llama3.2-3B / 24576 for DS-R1-Distill-Qwen-1.5B
Sampling temperature	1.0 for Qwen2.5 and Llama, 0.6 for DS-R1-Distill-Qwen-1.5B
Top-p	1.0

Table 3: **Training hyperparameters used in all GAPO experiments.** We use 8xH200 GPUs with verl (Sheng et al., 2025) framework for both training and evaluation. Training takes about 4-5 days for DS-R1 models and 1 day for Qwen2.5-Math-1.5B and Llama models.