NGRPO: Negative-Enhanced Group Relative Policy Optimization

Anonymous ACL submission

Abstract

Group-relative RLVR relies on reward variation among responses to the same prompt. When all sampled responses are incorrect, however, standard GRPO assigns zero advantages and discards a training signal that is common near the model's capability frontier. We call this failure mode the *Gradient Void*. We propose NGRPO, a statistics-only extension of GRPO for bounded verifiable rewards. NGRPO adds a virtual maximum-reward anchor only when computing group normalization statistics, turning homogeneous non-maximum groups into calibrated negative evidence while leaving real-sample ranking unchanged. Because this calibration induces a negative advantage shift, we introduce Dynamic Asymmetric Clipping to widen rare positive updates and conservatively bound failure-derived negative updates. Under matched training budgets, NGRPO improves Avg@64 on MATH500, AMC23, and AIME2025 across two base models, with the clearest gains on AIME2025; GPQA-Diamond and LiveCodeBench provide initial transfer evidence.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become a central tool for improving the reasoning ability of large language models (Cobbe et al., 2021; DeepSeek-AI, 2025; OpenAI, 2024; Lambert et al., 2024). In domains such as mathematics and programming, correctness can be checked automatically, allowing policy optimization to use outcome feedback without human preference labels or learned reward models. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is especially attractive in this setting: it removes the learned critic used in PPO (Schulman et al., 2017), samples multiple responses to the same prompt, and normalizes rewards within the group.

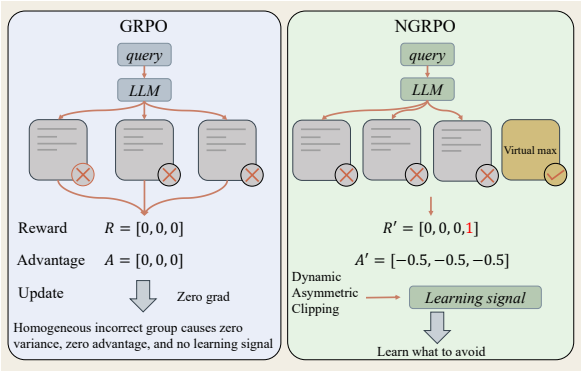


Figure 1: Overview of NGRPO. A virtual maximum-reward anchor is added only when computing group statistics. The anchor turns homogeneously incorrect groups from zero-gradient events into calibrated negative evidence, while Dynamic Asymmetric Clipping controls the strength of the resulting update.

The same normalization that stabilizes GRPO also creates a failure mode at the capability frontier. Group-relative learning needs contrast. If a prompt has both successful and failed rollouts, the successful trajectories receive positive advantages and the failed trajectories receive negative advantages. If every sampled response is wrong, however, all rewards are identical. The within-group variance vanishes, the normalized advantages collapse to zero, and the optimizer treats repeated failure as silence. We refer to this phenomenon as the *Gradient Void*.

This silence is undesirable. A homogeneously correct group and a homogeneously incorrect group are mathematically similar under standard normalization, but they have different learning semantics. An all-correct group is outcome-saturated: every sampled trajectory reaches the verifier's maximum reward, so the outcome signal provides little reason to prefer one sample over another. An all-wrong group is different: it records several failed searches around the same prompt. These failures show that the current sampled group did not find a verified

solution for the prompt, making additional exploration valuable.

Existing approaches expose a design tension. Filtering homogeneous groups, as in DAPO (Yu et al., 2025), is a stability-oriented choice that avoids degenerate updates, but it also removes hard prompts exactly when the model lacks successful rollouts. Negative-sample objectives make failures trainable by changing the reward or advantage assignment rule (Xiong et al., 2025; Zhu et al., 2025; Feng et al., 2025), while zero-variance and residual-prompt methods often introduce entropy guidance, history tracking, prompt replay, or process-like signals (Le et al., 2025; Liu et al., 2025a; Baroian and Berger, 2026; Yu et al., 2026). We ask whether all-wrong groups can be made useful within group-relative normalization itself, using only the bounded verifier structure already available in RLVR.

We propose NGRPO (Negative-Enhanced Group Relative Policy Optimization), a plug-in extension of GRPO for bounded verifiable rewards. As shown in Figure 1, NGRPO augments each reward group with one virtual rollout whose reward is the maximum attainable score. This virtual rollout is not generated by the model and never appears in the policy loss; it only anchors the mean and standard deviation used for advantage normalization. For all-wrong groups, the anchor restores non-zero variance and assigns every failed real rollout a finite negative advantage. For mixed groups, it preserves the ranking among real samples while shifting the baseline toward the verified optimum. For all-correct groups, the outcome signal remains saturated, so the method does not introduce artificial preference among already successful rollouts.

Because this calibration deliberately increases negative pressure, NGRPO also introduces Dynamic Asymmetric Clipping. The clipping range is adapted to the strength of the anchor-induced shift: when a group contains mostly failures, positive updates are allowed a wider trust region and negative updates are clipped more conservatively; when the group is already close to the maximum reward, the schedule moves toward a less exploratory clipping regime. This provides a bounded mechanism for increasing exploration pressure on failure-dominated groups.

Our contributions are:

- We identify the Gradient Void in group-relative RLVR: homogeneously incorrect groups provide no useful GRPO update even though they often mark the model's capability frontier.
- We introduce Advantage Calibration, which adds a virtual maximum-reward anchor to recover normalized negative evidence without changing the ranking among real samples.
- We analyze the anchor-induced negative advantage shift and derive a binary closed form that links update pressure to group difficulty.
- We propose Dynamic Asymmetric Clipping, a monotone control rule that stabilizes the added failure-derived pressure by decoupling positive and negative trust regions.
- We evaluate NGRPO under matched training budgets and a 64-sample math protocol. The gains are largest where successful rollouts are sparse, especially AIME2025, with GPQA-Diamond and LiveCodeBench serving as initial transfer checks.

2 Preliminaries

Group-relative policy optimization. Let q be a prompt sampled from a training distribution and let $\{o_i\}_{i=1}^{G}$ be G responses sampled from the behavior policy $\pi_{\theta_{\text{old}}}$. A verifier assigns each response an outcome reward r_i. Throughout the paper, rewards are bounded and normalized to $[0, r_{\max}]$, Std denotes the sample standard deviation with Bessel correction, and $\epsilon = 10^{-6}$ is used for numerical stability. GRPO estimates advantages by normalizing rewards within the prompt-level group:

$$\begin{aligned} A_i &= \frac{r_i - \mu_G}{\sigma_G + \epsilon}, \\ \mu_G &= \frac{1}{G}\sum_{j=1}^{G} r_j, \\ \sigma_G &= \mathrm{Std}(\{r_j\}_{j=1}^{G}). \end{aligned} \tag{1}$$

With the token-level importance ratio

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \tag{2}$$

the clipped group-relative objective is

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\big(\rho_{i,t}A_i, \mathrm{clip}(\rho_{i,t}, 1-\varepsilon, 1+\varepsilon)A_i\big)\Bigg]. \tag{3}$$

This objective is critic-free and adapts the update scale to prompt difficulty through the empirical reward spread in the group. A KL regularizer can be added as in PPO-style RLHF; we write the clipped surrogate because it is the common interface used by the group-relative variants compared in this paper.

Dynamic filtering and asymmetric clipping. DAPO improves large-scale RLVR by filtering groups that do not contain useful reward variation and by using decoupled clipping bounds (Yu et al., 2025). A group-level filter keeps optimization focused on prompts where the current sampler produces both better and worse responses. This is effective for stability, but it trades away homogeneous-failure groups. Decoupled clipping uses a larger upper bound and a smaller lower bound to make positive and negative updates obey different trust-region widths, a design we retain but make dynamic.

Negative reinforcement baselines. Another line of work increases the role of negative samples directly (Xiong et al., 2025; Zhu et al., 2025). A representative weighted REINFORCE-style formulation assigns fixed signed advantages,

$$A_i^{\text{W-R}} = \begin{cases} \alpha, & r_i = r_{\max}, \\ -1, & r_i < r_{\max}, \end{cases} \tag{4}$$

with a small positive weight α for successful samples. Such objectives guarantee that failed rollouts contribute gradients, including in all-wrong groups. They are best viewed as representative non-group-normalized negative-sample objectives: the update magnitude is fixed by the signed weighting rule rather than by the reward distribution within the prompt group. NGRPO targets the middle ground: it keeps normalized group-relative scaling while making collective failures trainable.

3 The Gradient Void

From zero variance to missing supervision. Consider a prompt whose sampled responses receive the same reward:

$$r_1 = \cdots = r_G = a. \tag{5}$$

Equation 1 gives $r_i - \mu_G = 0$ for every sample. The standard deviation is also zero, so the stabilizer prevents numerical overflow but cannot create contrast:

$$A_1 = \cdots = A_G = 0. \tag{6}$$

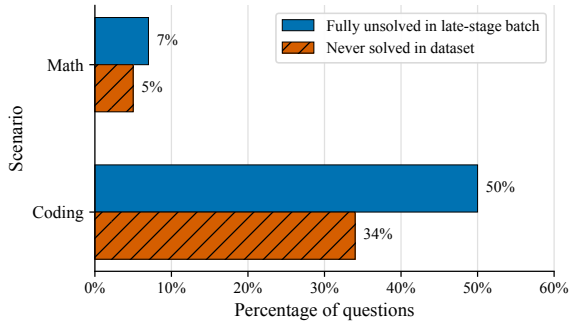


Figure 2: Homogeneously incorrect groups appear in both mathematical and coding RLVR. The phenomenon is stronger when the base model is weaker or the task is harder, suggesting that all-wrong groups concentrate near the current capability frontier.

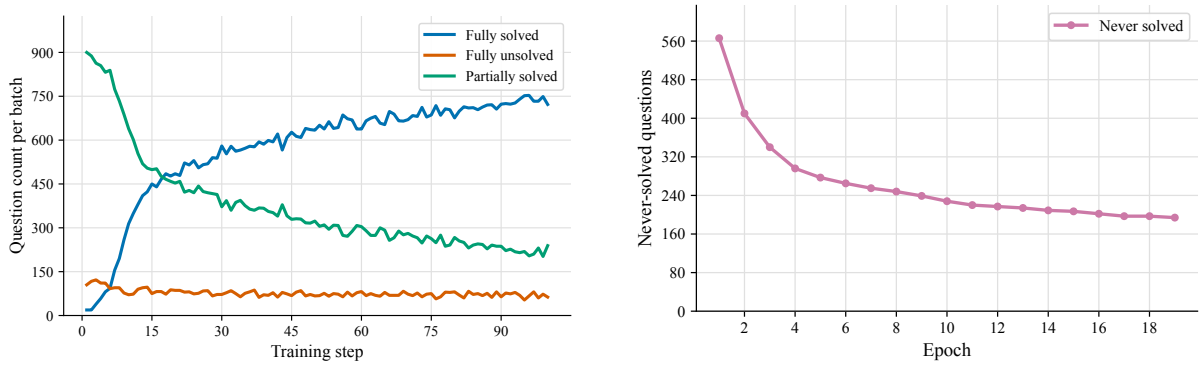
The group contributes no policy-gradient signal under standard group-relative normalization.

This zero update has different meanings at the two endpoints. If $a = r_{\max}$, every sampled response has already reached the verifier's maximum reward, so the outcome signal is saturated. If $a < r_{\max}$, the sampler has tried G trajectories for the same prompt and none reaches the verifier's maximum. This is not evidence for a correct alternative, but it is evidence that the sampled trajectories should not be reinforced as they are.

Empirical prevalence. The Gradient Void is not a rare corner case. Figure 2 shows that homogeneously incorrect groups appear in both math and coding RLVR, and become more frequent for weaker base models or harder tasks. Figure 3 shows the same effect during GRPO training on MATH: many partially solved groups become fully solved, but a persistent hard subset remains homogeneously incorrect. Filtering or zeroing these groups discards exactly the rollouts that expose where the current sampler still fails.

The scale is large enough to affect training rather than only diagnostics. In our prevalence analysis, about 7% of mathematical prompts and 50% of coding prompts never produce a correct rollout under the sampled budget. During training, homogeneously incorrect groups account for about 7% of math groups and 34% of coding groups. The higher rate in coding and the persistence of never-solved math prompts both match the mechanism: all-wrong groups concentrate where successful rollouts are sparse.

Design requirements. The objective should therefore satisfy three constraints. First, non-



(a) Per-batch group composition during training.

(b) Cumulative prompts that never produce a correct rollout.

Figure 3: The Gradient Void persists through training. Standard GRPO converts many partially solved groups into fully solved groups, but a hard subset remains homogeneously incorrect for most or all of training.

maximum homogeneous groups should produce non-zero negative advantages. Second, the signal should remain calibrated by the prompt-level group, preserving the stability benefits of GRPO and the ranking among real samples. Third, failure-derived updates should be conservative: an incorrect final answer identifies a failed trajectory, not the correct reasoning path, and may still contain useful intermediate steps. NGRPO implements these constraints with a virtual maximum-reward anchor and dynamic asymmetric clipping.

4 NGRPO

NGRPO modifies only the advantage computation and the clipping bounds of group-relative RLVR. It does not require a critic, a reward model, external traces, or additional generated rollouts. The core idea is to anchor local comparison to the known maximum reward of a bounded verifier.

4.1 Advantage Calibration

For each prompt-level reward group $R = (r_1, \ldots, r_G)$, NGRPO appends one virtual reward $r_{\max}$:

$$R^+ = (r_1, \ldots, r_G, r_{\max}). \tag{7}$$

The virtual sample is used only to compute normalization statistics. It is not a model output, has no tokens, and contributes no term to the policy loss. The calibrated advantage for each real response is

$$\begin{aligned} A_i^+ &= \frac{r_i - \mu^+}{\sigma^+ + \epsilon}, \\ \mu^+ &= \frac{\sum_{j=1}^{G} r_j + r_{\max}}{G+1}, \\ \sigma^+ &= \mathrm{Std}(R^+). \end{aligned} \tag{8}$$

This definition applies to binary correctness rewards and to bounded continuous rewards after clipping or normalizing them to $[0, r_{\max}]$, as long as a reliable upper bound $r_{\max}$ is known. Unless stated otherwise, we add one virtual anchor.

The anchor has different effects in different group regimes. For an all-correct group, all real rewards already equal $r_{\max}$, so the group remains outcome-saturated. For an all-wrong binary group, the augmented set contains G zeros and one positive reward; the real samples now receive finite negative advantages instead of zero. For a homogeneous but non-maximum continuous-reward group, every real sample is below the anchor and therefore receives a negative update. For a mixed group, the anchor shifts and rescales the real rewards but preserves their order:

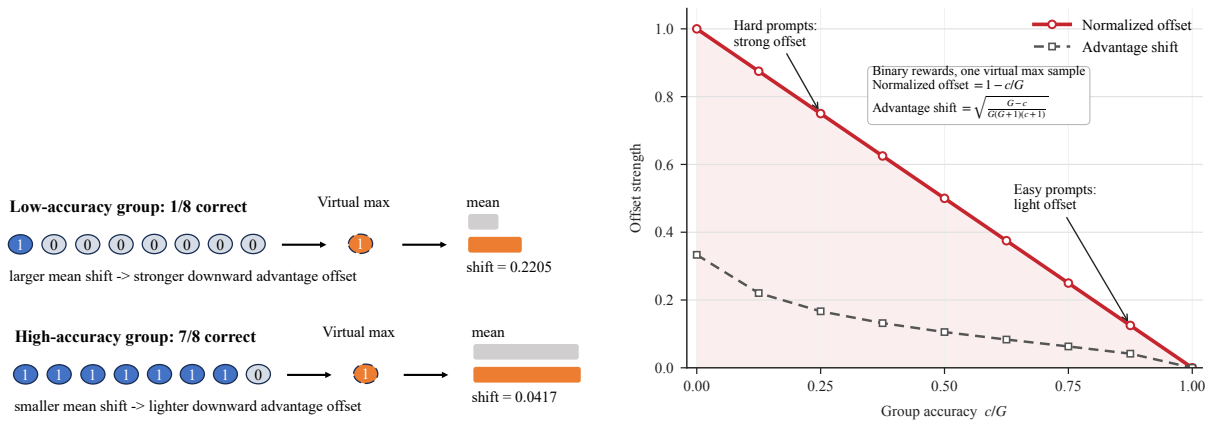
$$r_i > r_j \Rightarrow A_i^+ > A_j^+. \tag{9}$$

4.2 Binary Analysis and Advantage Shift

For binary rewards, the effect of the anchor has a closed form. Suppose $r_{\max} = 1$ and a group of size G contains c correct real responses and $G - c$ incorrect real responses. For $0 \le c < G$, the augmented group contains $c + 1$ positive rewards and $G - c$ zeros. Using the sample-standard-deviation convention defined in Section 2, the calibrated advantages are

$$\begin{aligned} A_{\text{correct}}^+ &= \sqrt{\frac{G(G-c)}{(G+1)(c+1)}}, \\ A_{\text{wrong}}^+ &= -\sqrt{\frac{G(c+1)}{(G+1)(G-c)}}. \end{aligned} \tag{10}$$

When $c = 0$, standard GRPO assigns every response zero advantage, whereas NGRPO assigns



(a) Calibrated advantages under representative group compositions.

(b) The negative shift grows as the group contains fewer successful rollouts.

Figure 4: Advantage Calibration induces a controlled negative shift. Groups with higher error rates receive stronger negative pressure, while groups already close to the maximum reward remain close to standard GRPO.

each failed response $-1/\sqrt{G+1}$. When $0 < c < G$, correct rollouts remain positive and wrong rollouts remain negative.

The anchor also shifts the average advantage over the real samples below zero:

$$\Delta_c = -\frac{1}{G}\sum_{i=1}^{G} A_i^+ = \sqrt{\frac{G-c}{G(G+1)(c+1)}}. \quad (11)$$

The shift is larger when fewer correct responses are present, and Figure 4 visualizes this monotonicity. Thus, NGRPO does not add an arbitrary penalty; it produces a group-calibrated negative signal whose magnitude reflects how far the sampled group is from the verifier's maximum.

4.3 Dynamic Asymmetric Clipping

Advantage Calibration deliberately increases negative pressure. This helps frontier prompts by reducing probability mass on sampled failed trajectories and encouraging alternative reasoning paths. It is also risky: a failed response says that one path did not work, but it does not identify the correct path. Negative updates should therefore be available but tightly bounded.

NGRPO adapts the upper and lower clipping radii according to the anchor-induced negative advantage shift. This schedule is a monotone control rule for update strength, not an optimal clipping solution. For binary rewards, the normalized shift score is

$$s_c = \sqrt{\frac{G-c}{G(c+1)}}. \quad (12)$$

This is the advantage shift in Equation 11 normalized by its all-wrong value. It is maximal when $c = 0$ and decreases to zero as the group approaches all-correct; for a binary group, we set $s_q = s_c$. For bounded continuous rewards, we use the same principle by computing the mean negative shift over real samples,

$$\begin{aligned} \Delta_q &= -\frac{1}{G}\sum_{i=1}^{G} A_i^+, \\ s_q &= \mathrm{clip}_{[0,1]}\left(\frac{\Delta_q}{1/\sqrt{G+1}+\epsilon}\right), \end{aligned} \quad (13)$$

where $1/\sqrt{G+1}$ is the all-wrong binary shift under the same sample-standard-deviation convention. In binary settings, Equation 13 reduces to Equation 12.

Dynamic Asymmetric Clipping sets

$$\varepsilon_q^+ = \varepsilon_{\min}^+ + s_q(\varepsilon_{\max}^+ - \varepsilon_{\min}^+), \quad (14)$$

$$\varepsilon_q^- = \varepsilon_{\max}^- - s_q(\varepsilon_{\max}^- - \varepsilon_{\min}^-). \quad (15)$$

The clipped ratio becomes

$$\bar{\rho}_{i,t}(\theta) = \mathrm{clip}\left(\rho_{i,t}(\theta), 1-\varepsilon_q^-, 1+\varepsilon_q^+\right). \quad (16)$$

When a group is dominated by failures, ε_q^+ increases and ε_q^- decreases: verified successes are rare and should be reinforced more strongly, whereas outcome-level negative updates should remain conservative because they do not identify the correct reasoning path. When a group is mostly correct, the schedule moves in the opposite direction, limiting over-reinforcement of redundant successes and allowing ordinary correction of lower-reward samples.

The final NGRPO objective is

$$\mathcal{J}_{\text{NGRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(\rho_{i,t}A_i^{+},\ \bar{\rho}_{i,t}A_i^{+}\right)\right]. \tag{17}$$

The method adds only $O(G)$ scalar operations per reward group and leaves the rollout, verifier, and model architecture unchanged.

5 Experiments

We evaluate whether NGRPO improves reasoning accuracy under a fixed rollout budget, whether the gains are larger on harder benchmarks, whether the method applies beyond mathematical training data, and whether its training dynamics match the intended controlled-exploration behavior.

5.1 Experimental Setup

Mathematical reasoning. Our main experiments fine-tune Qwen2.5-Math-7B (Yang et al., 2024) and Llama-3.1-8B-Instruct (AI at Meta, 2024) on the MATH training set (Hendrycks et al., 2021). For each base model, all math methods use a single 8-card NVIDIA H20 machine, learning rate 1×10^{-6}, global batch size 1024, group size $G = 8$, rollout temperature 1.0, and 20 training epochs. We evaluate in-domain mathematical reasoning on MATH500 (Hendrycks et al., 2021), AMC23 (math-ai, 2025), and AIME2025 (Test-Time Compute Organization, 2025), and use GPQA-Diamond (Rein et al., 2023) as an out-of-domain transfer check. For the three math benchmarks, each problem is sampled 64 times at temperature 0.6, which reduces sensitivity to stochastic decoding compared with single-sample evaluation. We report Avg@64, the mean correctness over sampled responses; Pass@64, whether at least one sampled response is correct; and Consistency@64, answer agreement among the sampled final answers for the same problem. We use Consistency@64 as a descriptive diagnostic of sampling concentration, not as a standalone quality metric. For GPQA-Diamond, we report Avg@4.

Baselines. For the main mathematical comparison in Table 1, we compare against PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), and W-REINFORCE (Xiong et al., 2025; Zhu et al., 2025). All methods within this comparison use the same training and evaluation budgets. For DAPO, we follow its asymmetric clipping setting with higher positive clipping. For W-REINFORCE, we set the positive advantage weight to 0.1. For NGRPO, Dynamic Asymmetric Clipping uses $\varepsilon^{+}_{\min} = 0.24$, $\varepsilon^{+}_{\max} = 0.28$, $\varepsilon^{-}_{\min} = 0.16$, and $\varepsilon^{-}_{\max} = 0.20$.

Coding. To test whether NGRPO can be used as a plug-in modification outside mathematics, we use a LiveCodeBench v5 coding setup (Jain et al., 2024). We fine-tune DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI, 2025) on the training split and compare GRPO with NGRPO under the same base configuration. This experiment is intended as a portability check rather than an exhaustive coding benchmark.

5.2 Main Mathematical Results

Table 1 shows that NGRPO improves Avg@64 in every evaluated math setting, although some absolute margins on MATH500 and AMC23 are small. This pattern is important for interpreting the results: the strongest evidence is not a uniform large gain across all benchmarks, but a larger gain where the method predicts it should help. On Qwen2.5-Math-7B, the hardest benchmark shows the clearest benefit: on AIME2025, NGRPO improves Avg@64 from the strongest baseline value of 9.95 to 10.57 and improves Pass@64 from 40.00 to 46.67. This supports the central hypothesis that failure-derived signal matters most when successful rollouts are sparse.

The Llama-3.1-8B-Instruct results show the same pattern under a less math-specialized base model. NGRPO improves MATH500 Avg@64 from 53.70 to 54.08 and Pass@64 from 81.60 to 84.00 relative to the strongest baselines. On AIME2025, all methods are near the floor in Avg@64, but NGRPO matches the best Pass@64 while retaining the best Avg@64. Consistency@64 is generally lower for NGRPO than for GRPO. This is consistent with broader sampling only when paired with Pass@64 gains; it is also a cost when a deployment setting depends on majority voting or a single consensus answer.

5.3 Out-of-Domain and Coding Results

Table 2 evaluates whether the benefit is specific to the MATH training distribution. The Qwen2.5-Math-7B model trained with NGRPO reaches 35.67 Avg@4 on GPQA-Diamond, outperforming GRPO

Methods	MATH500			AMC23			AIME2025		
	Avg@64	*Consistency @64*	*Pass@64*	*Avg@64*	*Consistency @64*	*Pass@64*	*Avg@64*	*Consistency @64*	*Pass@64*
Qwen2.5-Math-7B									
PPO	75.93	91.23	93.40	59.18	87.46	**95.00**	8.59	92.97	40.00
GRPO	76.28	93.38	91.80	61.37	90.66	90.00	9.84	94.32	40.00
DAPO	75.23	90.37	93.00	60.43	88.40	**95.00**	9.95	94.01	36.67
W-REINFORCE	75.86	91.20	**94.00**	61.80	90.00	**95.00**	9.64	93.80	40.00
NGRPO	**76.31**	90.61	**94.00**	**62.77**	87.30	**95.00**	**10.57**	92.86	**46.67**
Llama-3.1-8B-Instruct									
PPO	48.27	89.23	78.00	28.12	91.09	67.50	0.00	100.00	0.00
GRPO	52.40	91.71	77.80	25.94	91.48	67.50	0.00	100.00	0.00
DAPO	53.70	89.26	81.60	28.91	90.51	**72.50**	0.31	99.64	6.67
W-REINFORCE	49.84	91.34	77.40	26.48	92.42	65.00	0.21	99.79	**10.00**
NGRPO	**54.08**	89.40	**84.00**	**29.80**	87.97	**72.50**	**0.36**	99.69	**10.00**

Table 1: Mathematical reasoning results under the same training budget and 64-sample evaluation protocol. NGRPO improves Avg@64 in all settings, with the most meaningful gains on harder benchmarks where successful rollouts are sparse. Best results are bolded and second-best results are underlined.

LiveCodeBench v5		
DeepSeek-R1-Distill-Qwen-1.5B		
Methods	**Pass@1**	**Pass@2**
GRPO	16.85	21.15
NGRPO	**18.10(+1.25)**	**23.66(+2.51)**
GPQA-Diamond		
Qwen2.5-Math-7B		
Methods	**Avg@4**	
DAPO	33.71	
GRPO	34.53	
NGRPO	**35.67(+1.14)**	

Table 2: Out-of-domain and coding transfer checks. These results provide initial evidence beyond the main math setting: a model trained on mathematical data scores higher in this GPQA-Diamond Avg@4 evaluation under NGRPO, and NGRPO also improves LiveCodeBench v5 when applied as a plug-in replacement for GRPO.

and DAPO in this evaluation. In the coding setting, replacing GRPO with NGRPO improves LiveCodeBench v5 from 16.85 to 18.10 Pass@1 and from 21.15 to 23.66 Pass@2. These checks suggest that the mechanism is not obviously specific to the MATH verifier, although broader cross-domain evaluation is still needed.

5.4 Training Dynamics

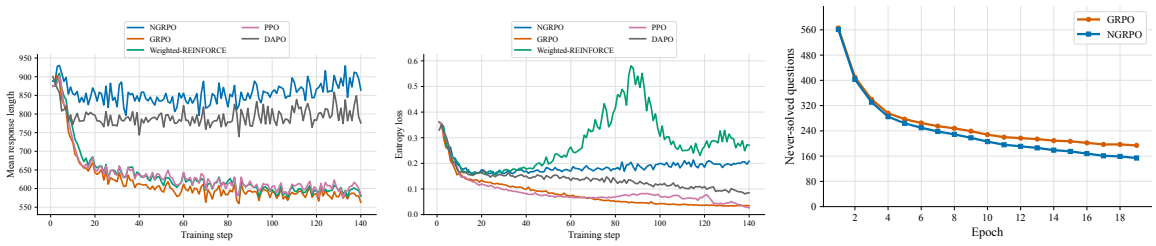
Figure 5 compares response length and entropy. PPO and GRPO show a tendency toward shorter responses and decreasing entropy, consistent with rapid narrowing of the policy distribution. Fixed-sign negative reinforcement keeps stronger exploration pressure but can produce less stable entropy behavior. NGRPO maintains a more controlled trajectory: response length remains moderate and entropy does not collapse early. Figure 5c provides a direct view of the intended mechanism: compared with GRPO, NGRPO leaves fewer prompts in the never-solved set by the end of training.

5.5 Ablation Study

Table 3 isolates the contributions of each component. "Advantage Calibration" toggles the virtual maximum-reward anchor in the normalization statistics; "Learn from All Failures" toggles whether homogeneously incorrect groups are retained rather than filtered; and "Dynamic Asymmetric Clipping" toggles the prompt-dependent clipping schedule in Equations 14–15. Advantage Calibration alone improves Pass@64 on both AIME2025 and AMC23, suggesting that the virtual anchor sharpens the signal within retained groups even when homogeneous failure groups are still filtered. Allowing the optimizer to learn from all-failure groups further raises Pass@64, but Avg@64 drops slightly, suggesting that the additional negative pressure can over-emphasize difficult prompts. Adding Dynamic Asymmetric Clipping mitigates this trade-off: the full method keeps the Pass@64 gains while raising Avg@64 to the best value on both benchmarks. This supports the interpretation that calibrated failure signal and conservative clipping are complementary rather than redundant.

6 Related Work

RLVR and group-relative optimization. Verifiable rewards make outcome supervision scalable for mathematical and programming reasoning (Cobbe et al., 2021; OpenAI, 2024; Lambert et al., 2024; DeepSeek-AI, 2025). GRPO removes PPO's learned critic by normalizing rewards across responses to the same prompt (Schulman et al., 2017;



(a) Response length. (b) Entropy loss. (c) Never-solved prompts.

Figure 5: Training dynamics on Qwen2.5-Math-7B. NGRPO maintains moderate response length, avoids early entropy collapse, and reduces the never-solved set.

Advantage Calibration	Learn from All Failures	Dynamic Asymmetric Clipping	AIME2025		AMC23	
			Avg@64	Pass@64	Avg@64	Pass@64
✗	✗	✗	9.84	40.00	61.37	90.00
✓	✗	✗	9.91	43.33	61.52	92.50
✓	✓	✗	9.78	**46.67**	61.24	**95.00**
✓	✓	✓	**10.57**	**46.67**	**62.77**	**95.00**

Table 3: Component ablation on Qwen2.5-Math-7B. Advantage Calibration, learning from all failures, and Dynamic Asymmetric Clipping contribute complementary gains.

Shao et al., 2024). Follow-up RLVR recipes improve sampling, filtering, clipping, normalization, or length bias, including DAPO, VAPO, REINFORCE++, and Dr. GRPO (Yu et al., 2025; Yue et al., 2025; Hu et al., 2025; Liu et al., 2025b). These works motivate our setting: prompt-level normalization is efficient, but it depends on within-group reward variation.

Zero-variance, residual, and negative groups. Recent methods also target zero-variance prompts, residual prompts, or negative groups, including RL-ZVP, ERPO, Prompt Replay, EP-GRPO, fixed-sign negative reinforcement, LENS, LUFFY, and ReLIFT (Le et al., 2025; Liu et al., 2025a; Baroian and Berger, 2026; Yu et al., 2026; Zhu et al., 2025; Feng et al., 2025; Yan et al., 2025; Ma et al., 2025). These approaches typically change the reward or advantage rule, add confidence or entropy guidance, track history, replay prompts, or introduce auxiliary training signals.

Positioning. NGRPO is complementary to these directions. It keeps prompt-level group-relative normalization and adds no entropy guidance, confidence model, history buffer, prompt replay, SFT trace, or extra rollout. Its only added signal is a statistic-only maximum-reward anchor for normalization statistics, which recovers calibrated negative evidence from all-wrong groups while preserving real-sample ranking; Dynamic Asymmetric Clipping bounds the induced negative shift.

7 Conclusion

We presented NGRPO, a group-relative RLVR extension that learns from collective failure without giving up normalized advantages. Homogeneously incorrect groups are zero-gradient events under standard GRPO, even though they mark prompts near the model's capability frontier. NGRPO restores this signal with a virtual maximum-reward anchor and controls the induced negative pressure with Dynamic Asymmetric Clipping. The results suggest that some frontier failures can be made trainable within group-relative normalization, while broader validation remains future work.

Limitations

NGRPO assumes a bounded verifier with a known maximum reward. This fits binary correctness and normalized partial-credit rewards, but open-ended, subjective, or noisy partial-credit rewards require reliable calibration or a different anchor. Our evidence is strongest for fixed-budget math training; some MATH500 and AMC23 margins are small, multiple seeds are still needed, and GPQA-Diamond and coding are transfer checks rather than exhaustive cross-domain evaluations. Finally, outcome-level negative evidence identifies failed trajectories, not which intermediate steps were useful, so clipping ranges should be retuned for new regimes.

References

AI at Meta. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Andrei Baroian and Rutger Berger. 2026. Prompt replay: Speeding up GRPO with on-policy reuse of high-signal prompts. *arXiv preprint arXiv:2603.21177*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yunzhen Feng, Parag Jain, Anthony Hartshorn, Yaqi Duan, and Julia Kempe. 2025. Don't waste mistakes: Leveraging negative RL-groups via confidence reweighting. *arXiv preprint arXiv:2510.08696*.

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. 2025. REINFORCE++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Thanh-Long V. Le, Myeongho Jeon, Kim Vu, Viet Lai, and Eunho Yang. 2025. No prompt left behind: Exploiting zero-variance prompts in LLM reinforcement learning via entropy-guided advantage shaping. *arXiv preprint arXiv:2509.21880*.

Chenxi Liu, Junjie Liang, Yuqi Jia, Bochuan Cao, Yang Bai, Heng Huang, and Xun Chen. 2025a. Explore data left behind in reinforcement learning for reasoning language models. *arXiv preprint arXiv:2511.04800*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, Bin Cui, and Wentao Zhang. 2025. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*.

math-ai. 2025. AMC23. `https://huggingface.co/datasets/math-ai/amc23`. Hugging Face dataset.

OpenAI. 2024. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*.

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, and Yang Wu. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Test-Time Compute Organization. 2025. AIME 2025: Unified test-time scaling format. `https://huggingface.co/datasets/test-time-compute/aime_2025`. Hugging Face dataset.

Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, and Hanze Dong. 2025. A minimalist approach to LLM reasoning: From rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and

16 others. 2025. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Song Yu, Li Li, Wenwen Zhao, and Zhisheng Yang. 2026. EP-GRPO: Entropy-progress aligned group relative policy optimization with implicit process guidance. *arXiv preprint arXiv:2605.04960*.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, Tiantian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, and 8 others. 2025. VAPO: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. The surprising effectiveness of negative reinforcement in LLM reasoning. *arXiv preprint arXiv:2506.01347*.

A Additional Sensitivity Analyses

A.1 Dynamic Asymmetric Clipping

Dynamic Asymmetric Clipping (DAC) has two roles: it widens positive updates when successful rollouts are sparse and tightens failure-derived negative updates when the anchor-induced shift is large. The default range used in the main experiments is $(\varepsilon^{+}_{\min}, \varepsilon^{+}_{\max}) = (0.24, 0.28)$ and $(\varepsilon^{-}_{\min}, \varepsilon^{-}_{\max}) = (0.16, 0.20)$. Table 4 shows the resulting bounds as the normalized shift score s_q varies. The full movement across the schedule is 0.04 on both sides, so DAC changes the clipping strength monotonically without introducing a large discontinuity.

s_q	ε^{+}_{q}	ε^{-}_{q}	Group regime
0.00	0.24	0.20	mostly solved
0.25	0.25	0.19	weak shift
0.50	0.26	0.18	moderate shift
0.75	0.27	0.17	strong shift
1.00	0.28	0.16	all-wrong binary

Table 4: DAC clipping bounds under the default schedule. Larger s_q corresponds to stronger anchor-induced negative shift.

The component ablation in Table 5 isolates the empirical effect of DAC under the matched-budget Qwen2.5-Math-7B setting. With Advantage Calibration and all-failure learning enabled, adding DAC improves Avg@64 on both AIME2025 and AMC23 while preserving the Pass@64 gains. This matches the intended role of DAC: it controls the extra negative pressure so that exploration gains do not reduce average correctness.

DAC	AIME Avg/Pass	AMC Avg/Pass
✗	9.78 / 46.67	61.24 / 95.00
✓	**10.57 / 46.67**	**62.77 / 95.00**

Table 5: Effect of DAC when Advantage Calibration and all-failure learning are enabled. Values are copied from the matched-budget component ablation in Table 3.

A.2 Number of Virtual Maximum-Reward Anchors

The main method adds one virtual maximum-reward anchor. More generally, let K denote the number of virtual anchors and define

$$R^{(K)} = (r_1, \ldots, r_G, \underbrace{r_{\max}, \ldots, r_{\max}}_{K \text{ anchors}}). \tag{18}$$

For binary rewards with $r_{\max} = 1$, group size G, and $c < G$ correct real responses, the augmented group contains $c + K$ positive rewards and $G - c$ zeros. Using the same sample-standard-deviation convention as the main text,

$$\begin{aligned} A^{(K)}_{\text{correct}} &= \sqrt{\frac{(G-c)(G+K-1)}{(G+K)(c+K)}}, \\ A^{(K)}_{\text{wrong}} &= -\sqrt{\frac{(c+K)(G+K-1)}{(G+K)(G-c)}}. \end{aligned} \tag{19}$$

The mean negative shift over real samples is

$$\Delta^{(K)}_{c} = \frac{K}{G}\sqrt{\frac{(G-c)(G+K-1)}{(G+K)(c+K)}}. \tag{20}$$

Thus, increasing K strengthens the negative shift and the magnitude of wrong-sample advantages. This can make all-wrong groups more influential, but it also increases the risk of over-penalizing failed trajectories that contain useful partial reasoning. For this reason, the main experiments use $K = 1$.

Table 6 shows the all-wrong binary case for the experimental group size $G = 8$. Standard GRPO corresponds to $K = 0$ and gives zero advantages in an all-wrong group. Adding one anchor recovers a finite negative signal; adding more anchors would make the negative update substantially stronger.

K	$A^{(K)}_{\text{wrong}}$	$\Delta^{(K)}_{0}$
0	0.000	0.000
1	-0.333	0.333
2	-0.474	0.474
4	-0.677	0.677

Table 6: Analytic effect of adding multiple virtual anchors for an all-wrong binary group with $G = 8$. Larger K increases failure-derived negative pressure.

Table 7 reports the matched-budget anchor-count comparison: the first two rows isolate one-anchor calibration, while the last two vary K under the full setting.

K	All-fail	DAC	AIME Avg/Pass	AMC Avg/Pass
0	✗	✗	9.84 / 40.00	61.37 / 90.00
1	✗	✗	9.91 / 43.33	61.52 / 92.50
1	✓	✓	**10.57 / 46.67**	**62.77 / 95.00**
2	✓	✓	10.16 / 46.67	61.74 / 92.50

Table 7: Empirical effect of evaluated anchor counts under matched-budget Qwen2.5-Math-7B ablations.

Under the full setting, increasing K from 1 to 2 lowers Avg@64 and AMC23 Pass@64, while $K > 4$ collapses midway through training, suggesting that excessive anchors over-amplify the negative shift in Eq. 20. We recommend $K \leq G/4$; for $G = 8$, we use $K = 1$.